```
DR   EMBL; AF133086; AAF00109.1; -.
DR   HSSP; P00763; 1DPO.
DR   MEROPS; S01.302; -.
DR   InterPro; IPR000859; CUB.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR002172; LDL_recept_A.
DR   InterPro; IPR001254; Trypsin.
DR   Pfam; PF00431; CUB; 2.
DR   Pfam; PF00057; ldl_recept_a; 4.
DR   Pfam; PF00089; trypsin; 1.
DR   PRINTS; PR00261; LDLRECEPTOR.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   SMART; SM00042; CUB; 2.
DR   SMART; SM00192; LDLa; 3.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS01180; CUB; 2.
DR   PROSITE; PS01209; LDLRA_1; 2.
DR   PROSITE; PS50068; LDLRA_2; 4.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Signal-anchor; Glycoprotein; Hydrolase; Serine protease;
KW   Transmembrane; Repeat.
FT   DOMAIN        1     55       CYTOPLASMIC (POTENTIAL).
FT   TRANSMEM     56     76       SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN)
FT                                (POTENTIAL).
FT   DOMAIN       77    855       EXTRACELLULAR (POTENTIAL).
FT   DOMAIN      214    334       CUB 1.
FT   DOMAIN      340    447       CUB 2.
FT   DOMAIN      452    487       LDL-RECEPTOR CLASS A 1.
FT   DOMAIN      487    524       LDL-RECEPTOR CLASS A 2.
FT   DOMAIN      524    560       LDL-RECEPTOR CLASS A 3.
FT   DOMAIN      566    603       LDL-RECEPTOR CLASS A 4.
FT   DOMAIN      615    854       SERINE PROTEASE.
FT   ACT_SITE    656    656       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    711    711       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    805    805       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   CARBOHYD    109    109       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    302    302       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    485    485       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    772    772       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   855 AA;  94760 MW;  26143132C01F99C9 CRC64;
```

```
  Query Match           99.9%;  Score 4676;  DB 1;  Length 855;
  Best Local Similarity 99.9%;  Pred. No. 0;
  Matches  854;  Conservative    0;  Mismatches    1;  Indels    0;  Gaps    0;

Qy       1 MGSDRARKGGGGPKDFGAGLKYNSRHEKVNGLEEGVEFLPVNNVKKVEKHGPGRWVVLAA 60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db       1 MGSDRARKGGGGPKDFGAGLKYNSRHEKVNGLEEGVEFLPVNNVKKVEKHGPGRWVVLAA 60

Qy      61 VLIGLLLVLLGIGFLVWHLQYRDVRVQKVFNGYMRITNENFVDAYENSNSTEFVSLASKV 120
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db      61 VLIGLLLVLLGIGFLVWHLQYRDVRVQKVFNGYMRITNENFVDAYENSNSTEFVSLASKV 120

Qy     121 KDALKLLYSGVPFLGPYHKESAVTAFSEGSVIAYYWSEFSIPQHLVEEAERVMAEERVVM 180
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     121 KDALKLLYSGVPFLGPYHKESAVTAFSEGSVIAYYWSEFSIPQHLVEEAERVMAEERVVM 180

Qy     181 LPPRARSLKSFVVTSVVAFPTDSKTVQRTQDNSCSFGLHARGVELMRFTTPGFPDSPYPA 240
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     181 LPPRARSLKSFVVTSVVAFPTDSKTVQRTQDNSCSFGLHARGVELMRFTTPGFPDSPYPA 240

Qy     241 HARCQWALRGDADSVLSLTFRSFDLASCDERGSDLVTVYNTLSPMEPHALVQLCGTYPPS 300
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     241 HARCQWALRGDADSVLSLTFRSFDLASCDERGSDLVTVYNTLSPMEPHALVQLCGTYPPS 300

Qy     301 YNLTFHSSQNVLLITLITNTERRHPGFEATFFQLPRMSSCGGRLRKAQGTFNSPYYPGHY 360
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     301 YNLTFHSSQNVLLITLITNTERRHPGFEATFFQLPRMSSCGGRLRKAQGTFNSPYYPGHY 360

       361 PPNIDCTWNIEVPNNQHVKVSFKFFYLLEPGVPAGTCPKDYVEINGEKYCGERSQFVVTS 420
```

```
           ||||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||
Db     361 PPNIDCTWNIEVPNNQHVKVRFKFFYLLEPGVPAGTCPKDYVEINGEKYCGERSQFVVTS 420

Qy     421 NSNKITVRFHSDQSYTDTGFLAEYLSYDSSDPCPGQFTCRTGRCIRKELRCDGWADCTDH 480
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     421 NSNKITVRFHSDQSYTDTGFLAEYLSYDSSDPCPGQFTCRTGRCIRKELRCDGWADCTDH 480

Qy     481 SDELNCSCDAGHQFTCKNKFCKPLFWVCDSVNDCGDNSDEQGCSCPAQTFRCSNGKCLSK 540
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     481 SDELNCSCDAGHQFTCKNKFCKPLFWVCDSVNDCGDNSDEQGCSCPAQTFRCSNGKCLSK 540

Qy     541 SQQCNGKDDCGDGSDEASCPKVNVVTCTKHTYRCLNGLCLSKGNPECDGKEDCSDGSDEK 600
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     541 SQQCNGKDDCGDGSDEASCPKVNVVTCTKHTYRCLNGLCLSKGNPECDGKEDCSDGSDEK 600

Qy     601 DCDCGLRSFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYID 660
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     601 DCDCGLRSFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYID 660

Qy     661 DRGFRYSDPTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKP 720
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     661 DRGFRYSDPTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKP 720

Qy     721 AEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLL 780
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     721 AEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLL 780

Qy     781 PQQITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPGVYT 840
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     781 PQQITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPGVYT 840

Qy     841 RLPLFRDWIKENTGV 855
           |||||||||||||||
Db     841 RLPLFRDWIKENTGV 855
```

RESULT

```
RESULT   1
ST14_HUMAN
ID   ST14_HUMAN     STANDARD;      PRT:   855 AA.
AC   Q9Y5Y6;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   SUPPRESSOR OF TUMORIGENICITY 14 (EC 3.4.21.-) (MATRIPTASE) (MEMBRANE-
DE   TYPE SERINE PROTEASE 1) (MT-SP1).
GN   ST14 OR PRSS14 OR SNC19.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=99303581; PubMed=10373424;
RA   Lin C.Y., Anders J., Johnson M., Sang Q.A., Dickson R.B.;
RT   "Molecular cloning of cDNA for matriptase, a matrix-degrading serine
RT   protease with trypsin-like activity.";
RL   J. Biol. Chem. 274:18231-18236(1999).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=99432178; PubMed=10500122;
RA   Takeuchi T., Shuman M.A., Craik C.S.;
RT   "Reverse biochemistry: Use of macromolecular protease inhibitors to
RT   dissect complex biological processes and identify a membrane-type
RT   serine protease in epithelial cancer and normal tissue.";
RL   Proc. Natl. Acad. Sci. U.S.A. 96:11054-11061(1999).
RN   [3]
RP   CHARACTERIZATION.
RC   TISSUE=Milk;
RX   MEDLINE=99303582; PubMed=10373425;
RA   Lin C.Y., Anders J., Johnson M., Dickson R.B.;
RT   "Purification and characterization of a complex containing matriptase
RT   and a Kunitz-type serine protease inhibitor from human milk.";
RL   J. Biol. Chem. 274:18237-18242(1999).
CC   -!- FUNCTION: DEGRADES EXTRACELLULAR MATRIX. PROPOSED TO PLAY A ROLE
CC       IN BREAST CANCER INVASION AND METASTASIS. EXHIBITS TRYPSIN-LIKE
CC       ACTIVITY AS DEFINED BY CLEAVAGE OF SYNTHETIC SUBSTRATES WITH ARG
CC       OR LYS AS THE P1 SITE.
CC   -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN (PROBABLE).
CC   -!- SIMILARITY: CONTAINS 4 LDL-RECEPTOR CLASS A DOMAINS.
CC   -!- SIMILARITY: CONTAINS 2 CUB DOMAINS.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF118224; AAD42765.2; -.
```

GenCore version 4.5

OM protein - protein search, using sw model

Run on: January 20, 2002, 09:34:49 ; Search time 14.81 Seconds
(without alignments)
2116.707 Million cell updates/sec

Title: US-09-421-213-2
Perfect score: 4681
Sequence: 1 MGSDRARKGGGGPKDFGAGL.....PGVYTRLPLFRDWIKENTGV 855

Scoring table: BLOSUM62
Gapop 10.0 , Gapext 0.5

Searched: 100059 seqs, 36664827 residues

Total number of hits satisfying chosen parameters: 100059

Minimum DB seq length: 0
Maximum DB seq length: 2000000000

Post-processing: Minimum Match 0%
Maximum Match 100%
Listing first 45 summaries

Database : SwissProt_39:*

Pred. No. is the number of results predicted by chance to have a score greater than or equal to the score of the result being printed, and is derived by analysis of the total score distribution.

SUMMARIES

Result No.	Score	Query Match	Length	DB	ID	Description
1	4676	99.9	855	1	ST14_HUMAN	Q9y5y6 homo sapien
2	3901	83.3	855	1	ST14_MOUSE	P56677 mus musculu
3	727	15.5	1034	1	ENTK_PIG	P98074 sus scrofa
4	712.5	15.2	1035	1	ENTK_BOVIN	P98072 bos taurus
5	692	14.8	1042	1	CORI_HUMAN	Q9y5q5 homo sapien
6	682.5	14.6	1019	1	ENTK_HUMAN	P98073 homo sapien
7	676.5	14.5	1069	1	ENTK_MOUSE	P97435 mus musculu
8	663.5	14.2	1113	1	CORI_MOUSE	Q9z319 mus musculu
9	600	12.8	490	1	TMS2_MOUSE	Q9jiq8 mus musculu
10	586.5	12.5	704	1	CRAR_MOUSE	P98064 mus musculu
11	577	12.3	699	1	CRAR_HUMAN	P48740 h complemen
12	560.5	12.0	492	1	TMS2_HUMAN	O15393 homo sapien
13	533.5	11.4	638	1	KAL_MOUSE	P26262 mus musculu
14	533	11.4	454	1	TMS3_HUMAN	P57727 homo sapien
15	518	11.1	603	1	CFAI_MOUSE	Q61129 mus musculu
16	518	11.1	604	1	CFAI_RAT	Q9wuw3 rattus norv
17	514.5	11.0	638	1	KAL_RAT	P14272 rattus norv
18	513	11.0	455	1	TMS5_MOUSE	Q9er04 mus musculu
19	509.5	10.9	790	1	PLMN_PIG	P06867 sus scrofa
20	501.5	10.7	786	1	STUB_DROME	Q05319 drosophila
21	500	10.7	638	1	KAL_HUMAN	P03952 homo sapien
22	496	10.6	416	1	HEPS_MOUSE	O35453 mus musculu
23	496	10.6	810	1	PLMN_HUMAN	P00747 homo sapien
24	495	10.6	457	1	TMS5_HUMAN	Q9h3s3 homo sapien
25	492.5	10.5	583	1	CFAI_HUMAN	P05156 homo sapien
26	491.5	10.5	812	1	PLMN_MOUSE	P20918 mus musculu
27	491	10.5	416	1	FA9_BOVIN	P00741 bos taurus
28	490.5	10.5	812	1	PLMN_BOVIN	P06868 bos taurus
29	490	10.5	417	1	HEPS_HUMAN	P05981 homo sapien
30	486	10.4	810	1	PLMN_MACMU	P12545 macaca mula
31	484.5	10.4	343	1	PSS8_HUMAN	Q16651 homo sapien
32	482	10.3	686	1	MAS2_HUMAN	O00187 homo sapien
33	480.5	10.3	416	1	HEPS_RAT	Q05511 rattus norv
34	480	10.3	625	1	FA11_HUMAN	P03951 homo sapien
35	479.5	10.2	461	1	FA9_HUMAN	P00740 homo sapien
36	479	10.2	343	1	PLMN_SHEEP	P81286 ovis aries
37	478	10.2	810	1	PLMN_ERIEU	Q29485 erinaceus e
38	477.5	10.2	342	1	PSS8_RAT	Q9es87 rattus norv
39	472.5	10.1	2616	1	NDL_DROME	P98159 drosophila
40	472	10.1	275	1	TRY3_ANOGA	P35037 anopheles g
41	471.5	10.1	452	1	FA9_CANFA	P19540 canis famil
42	467	10.0	275	1	TRY4_ANOGA	P35038 anopheles g
43	467	10.0	562	1	TPA_HUMAN	P00750 homo sapien
44	463	9.9	437	1	TMS4_HUMAN	Q9nrs4 homo sapien
45	461.5	9.9	267	1	TRY7_ANOGA	P35041 anopheles g

ALIGNMENTS

```
RESULT   1
ST14_HUMAN
ID   ST14_HUMAN     STANDARD;      PRT;   855 AA.
AC   Q9Y5Y6;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   SUPPRESSOR OF TUMORIGENICITY 14 (EC 3.4.21.-) (MATRIPTASE) (MEMBRANE-
DE   TYPE SERINE PROTEASE 1) (MT-SP1).
GN   ST14 OR PRSS14 OR SNC19.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=99303581; PubMed=10373424;
RA   Lin C.Y., Anders J., Johnson M., Sang Q.A., Dickson R.B.;
RT   "Molecular cloning of cDNA for matriptase, a matrix-degrading serine
RT   protease with trypsin-like activity.";
RL   J. Biol. Chem. 274:18231-18236(1999).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=99432178; PubMed=10500122;
RA   Takeuchi T., Shuman M.A., Craik C.S.;
RT   "Reverse biochemistry: Use of macromolecular protease inhibitors to
RT   dissect complex biological processes and identify a membrane-type
RT   serine protease in epithelial cancer and normal tissue.";
RL   Proc. Natl. Acad. Sci. U.S.A. 96:11054-11061(1999).
RN   [3]
RP   CHARACTERIZATION.
RC   TISSUE=Milk;
RX   MEDLINE=99303582; PubMed=10373425;
RA   Lin C.Y., Anders J., Johnson M., Dickson R.B.;
RT   "Purification and characterization of a complex containing matriptase
RT   and a Kunitz-type serine protease inhibitor from human milk.";
RL   J. Biol. Chem. 274:18237-18242(1999).
CC   -!- FUNCTION: DEGRADES EXTRACELLULAR MATRIX. PROPOSED TO PLAY A ROLE
CC       IN BREAST CANCER INVASION AND METASTASIS. EXHIBITS TRYPSIN-LIKE
CC       ACTIVITY AS DEFINED BY CLEAVAGE OF SYNTHETIC SUBSTRATES WITH ARG
CC       OR LYS AS THE P1 SITE.
CC   -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN (PROBABLE).
CC   -!- SIMILARITY: CONTAINS 4 LDL-RECEPTOR CLASS A DOMAINS.
CC   -!- SIMILARITY: CONTAINS 2 CUB DOMAINS.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF118224; AAD42765.2; -.
```

```
DR   EMBL; AF133086; AAF00109.1; -.
DR   HSSP; P00763; 1DPO.
DR   MEROPS; S01.302; -.
DR   InterPro; IPR000859; CUB.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR002172; LDL_recept_A.
DR   InterPro; IPR001254; Trypsin.
DR   Pfam; PF00431; CUB; 2.
DR   Pfam; PF00057; ldl_recept_a; 4.
DR   Pfam; PF00089; trypsin; 1.
DR   PRINTS; PR00261; LDLRECEPTOR.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   SMART; SM00042; CUB; 2.
DR   SMART; SM00192; LDLa; 3.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS01180; CUB; 2.
DR   PROSITE; PS01209; LDLRA_1; 2.
DR   PROSITE; PS50068; LDLRA_2; 4.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Signal-anchor; Glycoprotein; Hydrolase; Serine protease;
KW   Transmembrane; Repeat.
FT   DOMAIN       1     55       CYTOPLASMIC (POTENTIAL).
FT   TRANSMEM     56    76       SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN)
FT                               (POTENTIAL).
FT   DOMAIN       77    855      EXTRACELLULAR (POTENTIAL).
FT   DOMAIN       214   334      CUB 1.
FT   DOMAIN       340   447      CUB 2.
FT   DOMAIN       452   487      LDL-RECEPTOR CLASS A 1.
FT   DOMAIN       487   524      LDL-RECEPTOR CLASS A 2.
FT   DOMAIN       524   560      LDL-RECEPTOR CLASS A 3.
FT   DOMAIN       566   603      LDL-RECEPTOR CLASS A 4.
FT   DOMAIN       615   854      SERINE PROTEASE.
FT   ACT_SITE     656   656      CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE     711   711      CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE     805   805      CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   CARBOHYD     109   109      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     302   302      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     485   485      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     772   772      N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   855 AA;  94760 MW;  26143132C01F99C9 CRC64;

  Query Match           99.9%;  Score 4676;  DB 1;  Length 855;
  Best Local Similarity  99.9%;  Pred. No. 0;
  Matches  854;  Conservative    0;  Mismatches    1;  Indels    0;  Gaps    0;

Qy       1 MGSDRARKGGGGPKDFGAGLKYNSRHEKVNGLEEGVEFLPVNNVKKVEKHGPGRWVVLAA 60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db       1 MGSDRARKGGGGPKDFGAGLKYNSRHEKVNGLEEGVEFLPVNNVKKVEKHGPGRWVVLAA 60

Qy      61 VLIGLLLVLLGIGFLVWHLQYRDVRVQKVFNGYMRITNENFVDAYENSNSTEFVSLASKV 120
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db      61 VLIGLLLVLLGIGFLVWHLQYRDVRVQKVFNGYMRITNENFVDAYENSNSTEFVSLASKV 120

Qy     121 KDALKLLYSGVPFLGPYHKESAVTAFSEGSVIAYYWSEFSIPQHLVEEAERVMAEERVVM 180
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     121 KDALKLLYSGVPFLGPYHKESAVTAFSEGSVIAYYWSEFSIPQHLVEEAERVMAEERVVM 180

Qy     181 LPPRARSLKSFVVTSVVAFPTDSKTVQRTQDNSCSFGLHARGVELMRFTTPGFPDSPYPA 240
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     181 LPPRARSLKSFVVTSVVAFPTDSKTVQRTQDNSCSFGLHARGVELMRFTTPGFPDSPYPA 240

Qy     241 HARCQWALRGDADSVLSLTFRSFDLASCDERGSDLVTVYNTLSPMEPHALVQLCGTYPPS 300
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     241 HARCQWALRGDADSVLSLTFRSFDLASCDERGSDLVTVYNTLSPMEPHALVQLCGTYPPS 300

Qy     301 YNLTFHSSQNVLLITLITNTERRHPGFEATFFQLPRMSSCGGRLRKAQGTFNSPYYPGHY 360
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     301 YNLTFHSSQNVLLITLITNTERRHPGFEATFFQLPRMSSCGGRLRKAQGTFNSPYYPGHY 360

Qy     361 PPNIDCTWNIEVPNNQHVKVRFKFFYLLEPGVPAGTCPKDYVEINGEKYCGERSQFVVTS 420
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     361 PPNIDCTWNIEVPNNQHVKVRFKFFYLLEPGVPAGTCPKDYVEINGEKYCGERSQFVVTS 420

Qy     421 NSNKITVRFHSDQSYTDTGFLAEYLSYDSSDPCPGQFTCRTGRCIRKELRCDGWADCTDH 480
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     421 NSNKITVRFHSDQSYTDTGFLAEYLSYDSSDPCPGQFTCRTGRCIRKELRCDGWADCTDH 480

Qy     481 SDELNCSCDAGHQFTCKNKFCKPLFWVCDSVNDCGDNSDEQGCSCPAQTFRCSNGKCLSK 540
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     481 SDELNCSCDAGHQFTCKNKFCKPLFWVCDSVNDCGDNSDEQGCSCPAQTFRCSNGKCLSK 540

Qy     541 SQQCNGKDDCGDGSDEASCPKVNVVTCTKHTYRCLNGLCLSKGNPECDGKEDCSDGSDEK 600
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     541 SQQCNGKDDCGDGSDEASCPKVNVVTCTKHTYRCLNGLCLSKGNPECDGKEDCSDGSDEK 600

Qy     601 DCDCGLRSFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYID 660
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     601 DCDCGLRSFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYID 660

Qy     661 DRGFRYSDPTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKP 720
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     661 DRGFRYSDPTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKP 720

Qy     721 AEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLL 780
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     721 AEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLL 780

Qy     781 PQQITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPGVYT 840
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Db     781 PQQITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPGVYT 840

Qy     841 RLPLFRDWIKENTGV 855
           |||||||||||||||
Db     841 RLPLFRDWIKENTGV 855


RESULT   2
ST14_MOUSE
ID   ST14_MOUSE     STANDARD;      PRT;   855 AA.
AC   P56677;
DT   15-JUL-1999 (Rel. 38, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   SUPPRESSOR OF TUMORIGENICITY 14 (EC 3.4.21.-) (EPITHIN).
GN   ST14 OR PRSS14.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=C.B.17SCID; TISSUE=Thymus;
RX   MEDLINE=99216440; PubMed=10199918;
RA   Kim M.G., Chen C., Lyu M.S., Cho E.G., Park D., Kozak C.,
RA   Schwartz R.H.;
RT   "Cloning and chromosomal mapping of a gene isolated from thymic
RT   stromal cells encoding a new mouse type II membrane serine protease,
RT   epithin, containing four LDL receptor modules and two CUB domains.";
RL   Immunogenetics 49:420-428(1999).
RN   [2]
RP   REVISIONS TO 23; 321; 325; 343; 409-410 AND C-TERMINUS.
RC   STRAIN=C.B.17SCID; TISSUE=Thymus;
RA   Kim M.G., Chen C., Cho E.G., Park D., Schwartz R.H.;
RL   Submitted (MAR-2000) to the EMBL/GenBank/DDBJ databases.
RN   [3]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Breast tumor;
RA   Strausberg R.;
RL   Submitted (MAR-2001) to the EMBL/GenBank/DDBJ databases.
CC   -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN (PROBABLE).
CC   -!- TISSUE SPECIFICITY: HIGHLY EXPRESSED IN INTESTINE, KIDNEY, LUNG,
CC       AND THYMUS. NOT EXPRESSED IN SKELETAL MUSCLE, LIVER, HEART,
CC       TESTIS, AND BRAIN.
```

```
CC   -!- SIMILARITY: CONTAINS 4 LDL-RECEPTOR CLASS A DOMAINS.
CC   -!- SIMILARITY: CONTAINS 2 CUB DOMAINS.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF042822; AAD02230.3; -.
DR   EMBL; BC005496; AAH05496.1; -.
DR   MEROPS; S01.014; -.
DR   MGD; MGI:1338881; St14.
DR    nterPro; IPR000859; CUB.
DR    nterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR002172; LDL_recept_A.
DR   InterPro; IPR001254; Trypsin.
DR   Pfam; PF00431; CUB; 2.
DR   Pfam; PF00057; ldl_recept_a; 4.
DR   Pfam; PF00089; trypsin; 1.
DR   PRINTS; PR00261; LDLRECEPTOR.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   SMART; SM00042; CUB; 2.
DR   SMART; SM00192; LDLa; 3.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS01180; CUB; 2.
DR   PROSITE; PS01209; LDLRA_1; 2.
DR   PROSITE; PS50068; LDLRA_2; 4.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Signal-anchor; Glycoprotein; Hydrolase; Serine protease;
KW   Transmembrane; Repeat.
FT   DOMAIN        1     55       CYTOPLASMIC (POTENTIAL).
FT   TRANSMEM     56     76       SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN)
FT                                (POTENTIAL).
FT   DOMAIN       77    855       EXTRACELLULAR (POTENTIAL).
FT   DOMAIN      214    331       CUB 1.
FT   DOMAIN      340    444       CUB 2.
FT   DOMAIN      451    488       LDL-RECEPTOR CLASS A 1.
FT   DOMAIN      489    522       LDL-RECEPTOR CLASS A 2.
FT   DOMAIN      523    561       LDL-RECEPTOR CLASS A 3.
FT   DOMAIN      565    604       LDL-RECEPTOR CLASS A 4.
FT   DOMAIN      615    854       SERINE PROTEASE.
FT   ACT_SITE    656    656       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    711    711       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    805    805       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   CARBOHYD    107    107       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    302    302       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    365    365       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    421    421       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    489    489       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    772    772       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   855 AA;  94654 MW;  4F10E84DA2146DD5 CRC64;


  Query Match           83.3%;  Score 3901;  DB 1;  Length 855;
  Best Local Similarity  81.8%;  Pred. No. 4.2e-256;
  Matches  699;  Conservative   73;  Mismatches   83;  Indels    0;  Gaps    0;

Qy       1 MGSDRARKGGGGRKDFGAGLKYNSRHEKVNGLEEGVEFLPVNNVKKVEKHGPGRWVVLAA 60
           |||:| || ||| ||||||||||| | :|| |||||||| || ||||| || ||||| |
Db       1 MGSNRGRKAGGGSQDFGAGLKYNSRLENMNGFEEGVEFLPANNAKKVEKRGPRRWVVLVA 60

Qy      61 VLIGLLLVLLGIGFLVWHLQYRDVRVQKVFNGYMRITNENFVDAYENSNSTEFVSLASKV 120
           ||   ||: |  | ||||  ||:||||||||::|||||  |:|||||  ||||:|||||:|
Db      61 VLFSFLLLSLMAGLLVWHFHYRNVRVQKVFNGHLRITNEIFLDAYENSTSTEFISLASQV 120

Qy     121 KDALKLLYSGVPFLGPYHKESAVTAFSEGSVIAYYWSEFSIPQHLVEEAERVMAEERVVM 180
           |:|||||: || ||||||:|||||||||||||||||||| || || :| || ||||
Db     121 KEALKLLYNEVPVLGPYHKKSAVTAFSEGSVIAYYWSEFSIPPHLAEEVDRAMAVEKVVT 180

Qy     181 LPPRARSLKSFVVTSVVAFPTDSKTVQRTQDNSCSFGLHARGVELMRFTTPGFPDSPYPA 240
           ||||||:|||||:|||||| | : :|||||||||| ||| |  : ||||||||:|||||
Db     181 LPPRARALKSFVLTSVVAFPIDPRMLQRTQDNSCSFALHAHGAAVTRFTTPGFPNSPYPA 240

Qy     241 HARCQWALRGDADSVLSLTFRSFDLASCDERGSDLVTVYNTLSPMEPHALVQLCGTYPPS 300
           |||||| ||||||||||||||||:| ||| |||||||::|||||||:|:|||||: ||
Db     241 HARCQWVLRGDADSVLSLTFRSFDVAPCDEHGSDLVTVYDSLSPMEPHAVVRLCGTFSPS 300

Qy     301 YNLTFHSSQNVLLITLITNTERRHPGFEATFFQLPRMSSCGGRLRKAQGTFNSPYYPGHY 360
           ||||| ||||| |:||||||:||||||||||||||:|||||| |    ||||:||||||||
Db     301 YNLTFLSSQNVFLVTLITNTDRRHPGFEATFFQLPKMSSCGGFLSDTQGTFSSPYYPGHY 360

Qy     361 PPNIDCTWNIEVPNNQHVKVSFKFFYLLEPGVPAGTCPKDYVEINGEKYCGERSQFVVTS 420
           |||:|||||:||||::||| || |||::| || |:| |||||||||||||||||||||:|
Db     361 PPNINCTWNIKVPNNRNVKVRFKLFYLVDPNVPVGSCTKDYVEINGEKYCGERSQFVVSS 420

Qy     421 NSNKITVRFHSDQSYTDTGFLAEYLSYDSSDPCPGQFTCRTGRCIRKELRCDGWADCTDH 480
           ||:|||| |||| ||||||||||||||||:||||| | |:|||||||||||||||| |:
Db     421 NSSKITVHFHSDHSYTDTGFLAEYLSYDSNDPCPGMFMCKTGRCIRKELRCDGWADCPDY 480

Qy     481 SDELNCSCDAGHQFTCKNKFCKPLFWVCDSVNDCGDNSDEQGCSCPAQTFRCSNGKCLSK 540
           ||| | |:| ||||||:|||||||||||||||| |||:|||||| :|:||||||| :
Db     481 SDERYCRCNATHQFTCKNQFCKPLFWVCDSVNDCGDGSDEEGCSCPAGSFKCSNGKCLPQ 540

Qy     541 SQQCNGKDDCGDGSDEASCPKVNVVTCTKHTYRCLNGLCLSKGNPECDGKEDCSDGSDEK 600
           ||:|||||:|||||||||  |||||:|||:|||| |||||||||||||| |||||||||
Db     541 SQKCNGKDNCGDGSDEASCDSVNVVSCTKYTYRCQNGLCLSKGNPECDGKTDCSDGSDEK 600

Qy     601 DCDCGLRSFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYID 660
           :|||||||||:|||||||:|||||||||||||||||||:||||||||:||||||||: |
Db     601 NCDCGLRSFTKQARVVGGTNADEGEWPWQVSLHALGQGHLCGASLISPDWLVSAAHCFQD 660

Qy     661 DRGFRYSDPTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKP 720
           |: |:||| | ||||||| |||:||| |||| :|||||:|| |||||||||||||||||
Db     661 DKNFKYSDYTMWTAFLGLLDQSKRSASGVQELKLKRIITHPSFNDFTFDYDIALLELEKS 720

Qy     721 AEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLL 780
           |||::|||||||||:|||||||||||||||||: |||||||||||:|||||||||||:|:
Db     721 VEYSTVVRPICLPDATHVFPAGKAIWVTGWGHTKEGGTGALILQKGEIRVINQTTCEDLM 780

Qy     781 PQQITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPGVYT 840
           |||||||||||||||||||||||||||||| | |||:|||||||:|||||||||||||
Db     781 PQQITPRMMCVGFLSGGVDSCQGDSGGPLSSAEKDGRMFQAGVVSWGEGCAQRNKPGVYT 840

Qy     841 RLPLFRDWIKENTGV 855
           |||: ||||||:|||
Db     841 RLPVVRDWIKEHTGV 855


RESULT   3
ENTK_PIG
ID   ENTK_PIG       STANDARD;      PRT;  1034 AA.
AC   P98074;
DT   01-FEB-1996 (Rel. 33, Created)
DT   01-FEB-1996 (Rel. 33, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ENTEROPEPTIDASE PRECURSOR (EC 3.4.21.9) (ENTEROKINASE).
GN   PRSS7 OR ENTK.
OS   Sus scrofa (Pig).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Cetartiodactyla; Suina; Suidae; Sus.
OX   NCBI_TaxID=9823;
RN   [1]
RP   SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.
RC   TISSUE=Duodenal mucosa;
RX   MEDLINE=94327548; PubMed=8051081;
RA   Matsushima M., Ichinose M., Yahagi N., Kakei N., Tsukada S.,
RA   Miki K., Kurokawa K., Tashiro K., Shiokawa K., Shinomiya K.,
RA   Umeyama H., Inoue H., Takahashi T., Takahashi K.;
RT   "Structural characterization of porcine enteropeptidase.";
```

```
RL   J. Biol. Chem. 269:19976-19982(1994).
CC   -!- FUNCTION: RESPONSIBLE FOR INITIATING ACTIVATION OF PANCREATIC
CC       PROTEOLYTIC PROENZYMES (TRYPSIN, CHYMOTRYPSIN AND CARBOXYPEPTIDASE
CC       A). IT CATALYZES THE CONVERSION OF TRYPSINOGEN TO TRYPSIN WHICH IN
CC       TURN ACTIVATES OTHER PROENZYMES INCLUDING CHYMOTRYPSINOGEN,
CC       PROCARBOXYPEPTIDASES, AND PROELASTASES.
CC   -!- CATALYTIC ACTIVITY: SELECTIVE CLEAVAGE OF 6-LYS-|-ILE-7 BOND IN
CC       TRYPSINOGEN.
CC   -!- SUBUNIT: HETEROTRIMER OF A CATALYTIC (LIGHT) CHAIN, A MULTIDOMAIN
CC       (HEAVY) CHAIN, AND A MINI CHAIN.
CC   -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN (PROBABLE).
CC   -!- PTM: THE CHAINS ARE DERIVED FROM A SINGLE PRECURSOR THAT IS
CC       CLEAVED BY A TRYPSIN-LIKE PROTEASE.
CC   -!- PTM: THE MINI CHAIN MAY BE CLEAVED BY ELASTASE.
CC   -!- SIMILARITY: CONTAINS 2 LDL-RECEPTOR CLASS A DOMAINS.
CC   -!- SIMILARITY: CONTAINS 2 CUB DOMAINS.
CC   -!- SIMILARITY: CONTAINS 1 SEA DOMAIN.
CC   -!- SIMILARITY: CONTAINS 1 SRCR DOMAIN.
CC   -!- SIMILARITY: CONTAINS 1 MAM DOMAIN.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   ---------------------------------------------------------------------
CC     is SWISS-PROT entry is copyright. It is produced through a collaboration
CC    etween the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; D30799; BAA06459.1; -.
DR   HSSP; P00763; 1DPO.
DR   MEROPS; S01.156; -.
DR   InterPro; IPR000859; CUB.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR002172; LDL_recept_A.
DR   InterPro; IPR000998; MAM.
DR   InterPro; IPR000082; SEA.
DR   InterPro; IPR001190; SRCR.
DR   InterPro; IPR001254; Trypsin.
DR   Pfam; PF00431; CUB; 2.
DR   Pfam; PF00057; ldl_recept_a; 2.
DR   Pfam; PF00629; MAM; 1.
DR   Pfam; PF01390; SEA; 1.
DR   Pfam; PF00530; SRCR; 1.
DR   Pfam; PF00089; trypsin; 1.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   SMART; SM00042; CUB; 2.
DR   SMART; SM00192; LDLa; 2.
DR   SMART; SM00137; MAM; 1.
DR   SMART; SM00200; SEA; 1.
DR   SMART; SM00202; SR; 1.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS01180; CUB; 2.
DR   PROSITE; PS01209; LDLRA_1; 2.
DR   PROSITE; PS50068; LDLRA_2; 2.
DR   PROSITE; PS00740; MAM_1; 1.
DR   PROSITE; PS50060; MAM_2; 1.
DR   PROSITE; PS50024; SEA; 1.
DR   PROSITE; PS00420; SRCR_1; FALSE_NEG.
DR   PROSITE; PS50287; SRCR_2; 1.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Signal-anchor; Glycoprotein; Myristate; Hydrolase;
KW   Serine protease; Zymogen; Transmembrane; Repeat.
FT   CHAIN        52    117       NON-CATALYTIC M CHAIN (MINI CHAIN).
FT   CHAIN       118    799       NON-CATALYTIC H CHAIN (HEAVY CHAIN).
FT   CHAIN       800   1034       CATALYTIC L CHAIN (LIGHT CHAIN).
FT   TRANSMEM     19     47       SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN).
FT   DOMAIN       52    169       SEA.
FT   DOMAIN      197    238       LDL-RECEPTOR CLASS A 1.
FT   DOMAIN      240    349       CUB 1.
FT   DOMAIN      357    519       MAM.
FT   DOMAIN      539    649       CUB 2.
FT   DOMAIN      656    694       LDL-RECEPTOR CLASS A 2.
FT   DOMAIN      693    786       SRCR.
FT   DOMAIN      800   1034       SERINE PROTEASE.
FT   ACT_SITE    840    840       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    891    891       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    986    986       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   LIPID         2      2       MYRISTATE (POTENTIAL).
FT   DISULFID    199    212       BY SIMILARITY.
FT   DISULFID    206    225       BY SIMILARITY.
FT   DISULFID    219    236       BY SIMILARITY.
FT   DISULFID    658    670       BY SIMILARITY.
FT   DISULFID    665    683       BY SIMILARITY.
FT   DISULFID    677    692       BY SIMILARITY.
FT   DISULFID    787    911       INTERCHAIN (BY SIMILARITY).
FT   DISULFID    825    841       BY SIMILARITY.
FT   DISULFID    925    992       BY SIMILARITY.
FT   DISULFID    956    971       BY SIMILARITY.
FT   DISULFID    982   1010       BY SIMILARITY.
FT   CARBOHYD    116    116       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    147    147       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    170    170       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    194    194       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    283    283       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    343    343       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    350    350       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    403    403       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    455    455       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    485    485       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    518    518       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    549    549       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    645    645       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    697    697       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    701    701       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    721    721       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    740    740       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    761    761       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    804    804       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    863    863       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    902    902       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    964    964       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1034 AA;  114776 MW;  0388C64CF64CC368 CRC64;


  Query Match             15.5%;  Score 727;  DB 1;  Length 1034;
  Best Local Similarity   23.8%;  Pred. No. 1.4e-41;
  Matches  258;  Conservative  135;  Mismatches  330;  Indels  360;  Gaps   41;

Qy      55 WVVLAAVLIGLLLVLLGIGFLVWHLQYRDVRVQKVF--NGYMRIT-----NENFVDAYEN 107
           :  | |:|: |   |: : :|      :|  : |     | |:||     | |  |
Db      21 FTALFAILMVLCAGLIAVSWLTIKGSEKDAALGKSHEARGTMKITSGVTYNPNLQDKL-- 78

Qy     108 SNSTEFVSLASKVKDALKLLYSGVPFLGPYHKESAVTAFSEGSVIAYY------W-SEFS 160
             | :|  ||  ::  :  ::        | | |  |  ||||  :      | |: :
Db      79 --SVDFKVLAFDIQQMIGEIFQSSNLKNEY-KNSRVLQFENGSVIVIFDLLFAQWVSDEN 135

Qy     161 IPQHLVEEAERVMAEERVVM---------------------------------------- 180
           | : |::  |   : : |
Db     136 IKEELIQGIEANKSSQLVAFHIDVNSIDITESLENYSTTSPSTTSDKLTTSSPPATPGNV 195

Qy     181 ----LP---PRARSLKSFVV----TSVVAFPTDSKTVQRTQDNSCSFGLHARGVELMRFT 229
               ||   | | :||   |        :  |  |     :|       |  |:  :
Db     196 SIECLPGSRPCADALKCIAVDLFCDGELNCPDGSDEDSKICATACD------GKFLLTES 249

Qy     230 TPGFPDSPYP----AHARCQWALRGDADSVLSLTFRSFDLASCDER------GSDLV--- 276
           :  |  : ||    |  ||| :| :    : | |  |:  | |         ||  :
Db     250 SGSFDAAQYPKLSEASVVCQWIIRVNQGLSIELNFSYFNTYSMDVLNIYEGVGSSKILRA 309

Qy     277 ------------------------------------TVYNTLS--------------- 283
                                                | :|:
Db     310 SLWLMNPGTIRIFSNQVTVTFLIESDENDYIGFNATYTAFNSTELNNDEKINCNFEDGFC 369
```

```
Qy    284 ------------------------------------------------------------PM 285
                                                                       |:
Db    370 FWIQDLNDDNEWERIQGTTFPPFTGPNFDHTFGNASGFYISTPTGPGGRQERVGLLSLPL 429

Qy    286 EPHALVQLC----------------------------------GTYPPSYN---------- 302
          ||  |  :|                                  | |  ::|
Db    430 EP-TLEPVCLSFWYYMYGENVYKLSINISNDQNIEKIIFQKEGNYGENWNYGQVTLNETV 488

Qy    303 ---LTFHSSQNVLLITLITNTERRHPGF-------EATFF-----QLPRMSSCGG--RLR 345
              : |:: :|  | :: :        |       | |        :||  : |||   |
Db    489 EFKVAFNAFKNQFLSDIALDDISLTYGICNVSLYPEPTLVPTSPPELP--TDCGGPFELW 546

Qy    346 KAQGTFNSPYYPGHYPPNIDCTWNIEVPNNQHVKVSFKFFYLLEPGVPAGTCPKDYVEI- 404
          :   || |  :| :||  . | ||:    ::::: |: | |           | |||
Db    547 EPNTTFTSMNFPNNYPNQAFCVWNLNAQKGKNIQLHFEEFDLENIA--------DVVEIR 598

Qy    405 NGEK--------YCGERSQFVVTSNSNKITVRFHSDQSYTDTGFLAEYLS---YDSSDPC 453
          :||:        | |       | | :|::|| | :: : |  || | : :       :||
Db    599 DGEEDDSLLLAVYTGPGPVEDVFSTTNRMTVLFITNDALTKGGFKANFTTGYHLGIPEPC 658

Qy    454 -PGQFTCRTGRCIRKELRCDGWADCTDHSDELNCSCDAGHQFTCKNKFCKPLFWVCDSVN 512
           | |  | |:    |||:: | :| ||| :|          :|          :|
Db    659 KEDNFQCENGECVLLVNLCNGFSHCKDGSDEAHCV----------RF----------LN 697

Qy    513 DCGDNSDEQGCSCPAQTFRLS---NGKCLSKSQQCNGKDDC---GDGSDEASCP------ 560
             :||          ||     :  |          | |   | |:  :| |
Db    698 GTANNSG-------LVQFRYSIWHTACAENWTTQTSDDVCQLLGLGTGNSSMPFFSSGG 750

Qy    561 ----KVNVVTCTKHTYRCLNGLCLSKGNPECDGKED------CSDGSDEKDCDCGLRSFT 610
              |:|         ||  :   : :|   ||      |:  |       || :
Db    751 GPFVKLNTAP---------NGSLILTASEQC--FEDSLILLQCNHKS------CGKKQVA 793

Qy    611 RQA--RVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYIDDRGFRYSD 668
          ::   ::||| |: || ||| |:|:  ||  :|||||:| :||||||||       |  :
Db    794 QEVSPKIVGGNDSREGAWPWVVALYYNGQ-LLCGASLVSRDWLVSAAHCVYG----RNLE 848

Qy    669 PTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKPAEYSSMVR 728
          |::| | ||||  |  ::| :. | :  |: :| :|    | |||:: ||    |:  ::
Db    849 PSKWKAILGLHMTSNLTSPQIVTRLIDEIVINPHYNRRRKDSDIAMMHLEFKVNYTDYIQ 908

Qy    729 PICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLLPQ-QITPR 787
          |||||: : ||| |:   : |||   | |: | |||: :: :::   |:  :|:  ||
Db    909 PICLPEENQVFPPGRICSIAGWGKVIYQGSPADILQEADVPLLSNEKCQQQMPEYNITEN 968

Qy    788 MMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPGVYTRLPLFRD 847
          ||| |:  ||:|||||||||||  :|.: |   ||| |:|  ||  |:|||| |:| | :
Db    969 MMCAGYEEGGIDSCQGDSGGPLMCLE-NNRWLLAGVTSFGYQCALPNRPGVYARVPKFTE 1027

Qy    848 WIK 850
          ||:
Db   1028 WIQ 1030


RESULT   4
ENTK_BOVIN
ID   ENTK_BOVIN     STANDARD;      PRT;  1035 AA.
AC   P98072;
DT   01-FEB-1996 (Rel. 33, Created)
DT   01-FEB-1996 (Rel. 33, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ENTEROPEPTIDASE PRECURSOR (EC 3.4.21.9) (ENTEROKINASE).
GN   PRSS7 OR ENTK.
OS   Bos taurus (Bovine).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Cetartiodactyla; Ruminantia; Pecora; Bovoidea;
OC   Bovidae; Bovinae; Bos.
OX   NCBI_TaxID=9913;
RN   [1]
RP   SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.
RC   TISSUE=Duodenum;
RX   MEDLINE=94329561; PubMed=8052624;
RA   Kitamoto Y., Yuan X., Wu Q., McCourt D.W., Sadler J.E.;
RT   "Enterokinase, the initiator of intestinal digestion, is a mosaic
RT   protease composed of a distinctive assortment of domains.";
RL   Proc. Natl. Acad. Sci. U.S.A. 91:7588-7592(1994).
RN   [2]
RP   SEQUENCE OF 801-1035 FROM N.A., AND PARTIAL SEQUENCE.
RX   MEDLINE=94043122; PubMed=8226855;
RA   Lavallie E.R., Rehemtulla A., Racie L.A., Diblasio E.A.,
RA   Ferenz C., Grant K.L., Light A., McCoy J.M.;
RT   "Cloning and functional expression of a cDNA encoding the catalytic
RT   subunit of bovine enterokinase.";
RL   J. Biol. Chem. 268:23311-23317(1993).
RN   [3]
RP   SEQUENCE OF 801-827.
RC   TISSUE=Intestine;
RX   MEDLINE=92189715; PubMed=1799406;
RA   Light A., Janska H.;
RT   "The amino-terminal sequence of the catalytic subunit of bovine
RT   enterokinase.";
RL   J. Protein Chem. 10:475-480(1991).
CC   -!- FUNCTION: RESPONSIBLE FOR INITIATING ACTIVATION OF PANCREATIC
CC       PROTEOLYTIC PROENZYMES (TRYPSIN, CHYMOTRYPSIN AND CARBOXYPEPTIDASE
CC       A). IT CATALYZES THE CONVERSION OF TRYPSINOGEN TO TRYPSIN WHICH IN
CC       TURN ACTIVATES OTHER PROENZYMES INCLUDING CHYMOTRYPSINOGEN,
CC       PROCARBOXYPEPTIDASES, AND PROELASTASES.
CC   -!- CATALYTIC ACTIVITY: SELECTIVE CLEAVAGE OF 6-LYS-|-ILE-7 BOND IN
CC       TRYPSINOGEN.
CC   -!- SUBUNIT: HETERODIMER OF A CATALYTIC (LIGHT) CHAIN AND A
CC       MULTIDOMAIN (HEAVY) CHAIN LINKED BY A DISULFIDE BOND.
CC   -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN (PROBABLE).
CC   -!- TISSUE SPECIFICITY: INTESTINAL BRUSH BORDER.
CC   -!- PTM: THE CHAINS ARE DERIVED FROM A SINGLE PRECURSOR THAT IS
CC       CLEAVED BY A TRYPSIN-LIKE PROTEASE.
CC   -!- SIMILARITY: CONTAINS 2 LDL-RECEPTOR CLASS A DOMAINS.
CC   -!- SIMILARITY: CONTAINS 2 CUB DOMAINS.
CC   -!- SIMILARITY: CONTAINS 1 SEA DOMAIN.
CC   -!- SIMILARITY: CONTAINS 1 SRCR DOMAIN.
CC   -!- SIMILARITY: CONTAINS 1 MAM DOMAIN.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; U09859; AAB40026.1; -.
DR   EMBL; L19663; AAA16035.1; -.
DR   PIR; A61436; A61436.
DR   HSSP; P00763; 1DPO.
DR   MEROPS; S01.156; -.
DR   InterPro; IPR000859; CUB.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR002172; LDL_recept_A.
DR   InterPro; IPR000998; MAM.
DR   InterPro; IPR000082; SEA.
DR   InterPro; IPR001190; SRCR.
DR   InterPro; IPR001254; Trypsin.
DR   Pfam; PF00431; CUB; 2.
DR   Pfam; PF00057; ldl_recept_a; 2.
DR   Pfam; PF00629; MAM; 1.
DR   Pfam; PF01390; SEA; 1.
DR   Pfam; PF00530; SRCR; 1.
DR   Pfam; PF00089; trypsin; 1.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   SMART; SM00042; CUB; 2.
DR   SMART; SM00192; LDLa; 2.
DR   SMART; SM00137; MAM; 1.
DR   SMART; SM00200; SEA; 1.
DR   SMART; SM00202; SR; 1.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS01180; CUB; 2.
```

```
DR   PROSITE; PS01209; LDLRA_1; 2.
DR   PROSITE; PS50068; LDLRA_2; 2.
DR   PROSITE; PS00740; MAM_1; 1.
DR   PROSITE; PS50060; MAM_2; 1.
DR   PROSITE; PS50024; SEA; 1.
DR   PROSITE; PS00420; SRCR_1; FALSE_NEG.
DR   PROSITE; PS50287; SRCR_2; 1.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Signal-anchor; Glycoprotein; Myristate; Hydrolase;
KW   Serine protease; Zymogen; Transmembrane; Repeat; Alternative splicing.
FT   CHAIN         1    800       NON-CATALYTIC CHAIN (HEAVY CHAIN).
FT   CHAIN       801   1035       CATALYTIC CHAIN (LIGHT CHAIN).
FT   TRANSMEM     19     47       SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN).
FT   DOMAIN       54    169       SEA.
FT   DOMAIN      197    238       LDL-RECEPTOR CLASS A 1.
FT   DOMAIN      240    350       CUB 1.
FT   DOMAIN      358    520       MAM.
FT   DOMAIN      540    650       CUB 2.
FT   DOMAIN      657    695       LDL-RECEPTOR CLASS A 2.
FT   DOMAIN      694    787       SRCR.
FT   MAIN        801   1035       SERINE PROTEASE.
FT   ACT_SITE    841    841       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    892    892       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    987    987       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   LIPID         2      2       MYRISTATE (POTENTIAL).
FT   DISULFID    199    212       BY SIMILARITY.
FT   DISULFID    206    225       BY SIMILARITY.
FT   DISULFID    219    236       BY SIMILARITY.
FT   DISULFID    659    671       BY SIMILARITY.
FT   DISULFID    666    684       BY SIMILARITY.
FT   DISULFID    678    693       BY SIMILARITY.
FT   DISULFID    788    912       INTERCHAIN (BY SIMILARITY).
FT   DISULFID    826    842       BY SIMILARITY.
FT   DISULFID    926    993       BY SIMILARITY.
FT   DISULFID    957    972       BY SIMILARITY.
FT   DISULFID    983   1011       BY SIMILARITY.
FT   CARBOHYD    116    116       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    147    147       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    170    170       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    194    194       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    233    233       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    263    263       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    264    264       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    404    404       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    456    456       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    486    486       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    519    519       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    550    550       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    646    646       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    698    698       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    722    722       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    741    741       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    762    762       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    864    864       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    903    903       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    965    965       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   VARSPLIC    166    192       MISSING (IN SHORT ISOFORM).
FT   CONFLICT    808    808       R -> Y (IN REF. 3).
SQ   SEQUENCE   1035 AA;  114887 MW;  E207970B08296E13 CRC64;


  Query Match           15.2%;  Score 712.5;  DB 1;  Length 1035;
  Best Local Similarity   23.3%;  Pred. No. 1.3e-40;
  Matches  253;  Conservative  149;  Mismatches  319;  Indels  365;  Gaps   45;

Qy      55 WVVLAAVLIGLLLVLLGIGFLVWHLQYRDVRVQKVF--NGYMR-ITNENFVDAYENSNST 111
           : ||  :|: |   |: : :|      :|    |     | :: |:   :    ::  |
Db      21 FAVLFVILVALCAGLIAVSWLSIQGSVKDAAFGKSHEARGTLKIISGATYNPHLQDKLSV 80

Qy     112 EFVSLASKVKDALKLLYSGVPFLGPYHKESAVTAFSEGSVIAYY------W-SEFSIPQH 164
           :|  ||  ::  :  ::    .  | | | |  |  ||:|  :     | |: :: :

Db      81 DFKVLAFDIQQMIDDIFQSSNLKNEY-KNSRVLQFENGSIIVIFDLLFDQWVSDKNVKEE 139

Qy     165 LVEEAERVMAEERVVM-----------------LPPRARSLKSFVVTSV---------V 197
           |:: |   : : |                    : |   | |  : ||:         :
Db     140 LIQGIEANKSSQLVTFHIDLNSIDITASLENFSTISPATTSEK--LTTSIPLATPGNVSI 197

Qy     198 AFPTDSKTVQRTQDNSCSFGLHARGVELM----------------------RFTTPGF 233
             | ||:         |:  |    ::|                        ||   |
Db     198 ECPPDSRL--------CADALKCIAIDLFCDGELNCPDGSDEDNKTCATACDGRFLLTGS 249

Qy     234 PDS------PYPAH---ARCQWALRGDADSVLSLTFRSFD------------------- 264
           |      | |::   | |:| :| :    : | |  |:
Db     250 SGSFEALHYPKPSNNTSAVCRWIIRVNQGLSIQLNFDYFNTYYADVLNIYEGMGSSKILR 309

Qy     265 ----------------------LASCDERGSDLV---TVYNTLSPMEPHALVQL------ 293
                                |  ||  ||: :    |   :  | :   ::
Db     310 ASLWSNNPGIIRIFSNQVTATFLIQSDE--SDYIGFKVTYTAFNSKELNNYEKINCNFED 367

Qy     294 --C-----------------GTYPPSYNLTF-HS-------------------------- 307
           |                  |:|||   || |:
Db     368 GFCFWIQDLNDDNEWERTQGSTFPPSTGPTFDHTFGNESGFYISTPTGPGGRRERVGLLT 427

Qy     308 -------------------SQNVLLITLITNTERRHPGFEATFFQ--------------- 333
                               :||  :::  ::::     | | ||
Db     428 LPLDPTPEQACLSFWYYMYGENVYKLSINISSDQ---NMEKTIFQKEGNYGQNWNYGQVT 484

Qy     334 ------------------------------------------------LPRMSSCGG 342
                                                           ||  : |||
Db     485 LNETVEFKVSFYGFKNQILSDIALDDISLTYGICNVSVYPEPTLVPTPPPELP--TDCGG 542

Qy     343 --RLRKAQGTFNSPYYPGHYPPNIDCTWNIEVPNNQHVKVSFKFFYLLEPGVPAGTCPKD 400
             | :   || |  :|  ||    | ||:     ::::: |: | |            |
Db     543 PHDLWEPNTTFTSINFPNSYPNQAFCIWNLNAQKGKNIQLHFQEFDLENIA--------D 594

Qy     401 YVEI-NGEKYCGERSQFV----------VTSNSNKITVRFHSDQSYTDTGFLAEYLS-- 446
           |||  :||   |: | | |:          | | :|::|| | :|      ||:| : :
Db     595 VVEIRDGE---GDDSLFLAVYTGPGPVNDVFSTTNRMTVLFITDNMLAKQGFKANFTTGY 651

Qy     447 -YDSSDPC-PGQFTCRTGRCIRKELRCDGWADCTDHSDELNC------SCDAGH--QFTC 496
              :||    | |: | ||     |||:  | | ||| :|      : |:    ||
Db     652 GLGIPEPCKEDNFQCKDGECIPLVNLCDGFPHCKDGSDEAHCVRLFNGTTDSSGLVQFRI 711

Qy     497 KNKFCKPLFWVCDSVNDCGDNSDEQGCSCPAQTFRCSNGKCLSKSQQCNGKDDCGDGSDE 556
           ::       |       | :|          | : |:  |     |  |    | |:
Db     712 QS------IWHV----ACAEN----------WTTQISDDVC-----QLLG---LGTGNSS 743

Qy     557 ASC------PKVNVVTCTKHTYRCLNGLCLSKGNPECDGKEDCSDGSDEKDCD---CGLR 607
                   | ||: |          ||  :   : :|       |      |:   || :
Db     744 VPTFSTGGGPYVNLNTAP-------NGSLILTPSQQC-----LEDSLILLQCNYKSCGKK 791

Qy     608 SFTRQA--RVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYIDDRGFR 665
           |::   ::|||:|: || ||| |:|:   |  :|||||:| :|||||||||       |
Db     792 LVTQEVSPKIVGGSDSREGAWPWVVALYFDDQ-QVCGASLVSRDWLVSAAHCVYG----R 846

Qy     666 YSDPTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKPAEYSS 725
           :|::| | ||||  |  ::| :: | : :|: :| :|    : |||:: ||    |:
Db     847 NMEPSKWKAVLGLHMASNLTSPQIETRLIDQIVINPHYNKRRKNNDIAMMHLEMKVNYTD 906

Qy     726 MVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLLPQ-QI 784
           ::|||||: : ||| |:   : |||   | |: | :||: :: :::   |:  :|:  |
Db     907 YIQPICLPEENQVFPPGRICSIAGWGALIYQGSTADVLQEADVPLLSNEKCQQQMPEYNI 966

Qy     785 TPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPGVYTRLPL 844
           |. |:| |: :|||||||||||||||   | : :|   ||| |:|  ||  |:|||| |:|
Db     967 TENMVCAGYEAGGVDSCQGDSGGPLMCQE-NNRWLLAGVTSFGYQCALPNRPGVYARVPR 1025

Qy     845 FRDWIK 850
           | :||:
Db    1026 FTEWIQ 1031


RESULT   5
CORI_HUMAN
```

```
ID   CORI_HUMAN     STANDARD;      PRT;  1042 AA.
AC   Q9Y5Q5; Q9UHY2;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ATRIAL NATRIURETIC PEPTIDE-CONVERTING ENZYME (EC 3.4.21.-) (PRO-ANP-
DE   CONVERTING ENZYME) (CORIN) (HEART SPECIFIC SERINE PROTEINASE ATC2).
GN   CRN.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Heart;
RX   MEDLINE=99262646; PubMed=10329693;
RA   Yan W., Sheng N., Seto M., Morser J., Wu Q.;
RT   "Corin, a mosaic transmembrane serine protease encoded by a novel cDNA
RT   from human heart.";
RL   J. Biol. Chem. 274:14926-14935(1999).
RN   [2]
RP   SEQUENCE OF 734-1040 FROM N.A.
RC   TISSUE=Heart;
RX   MEDLINE=20534769; PubMed=11082206;
RA   Hooper J.D., Scarman A.L., Clarke B.E., Normyle J.F., Antalis T.M.;
RT   "Localization of the mosaic transmembrane serine protease corin to
RT   heart myocytes.";
RL   Eur. J. Biochem. 267:6931-6937(2000).
RN   [3]
RP   CHARACTERIZATION.
RX   MEDLINE=20359740; PubMed=10880574;
RA   Yan W., Wu F., Morser J., Wu Q.;
RT   "Corin, a transmembrane cardiac serine protease, acts as a pro-atrial
RT   natriuretic peptide-converting enzyme.";
RL   Proc. Natl. Acad. Sci. U.S.A. 97:8525-8529(2000).
CC   -!- FUNCTION: CONVERTS PRO-ANP TO ANP. CLEAVES PRO-ANP SPECIFICALLY
CC       BETWEEN ARG-123 AND SER-124.
CC   -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN.
CC   -!- TISSUE SPECIFICITY: HIGHLY EXPRESSED IN HEART. EXPRESSED IN HEART
CC       MYOCYTES.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   -!- SIMILARITY: CONTAINS 7 LDL-RECEPTOR CLASS A DOMAINS.
CC   -!- SIMILARITY: CONTAINS 2 FRIZZLED (FZ) DOMAINS.
CC   -!- SIMILARITY: CONTAINS 1 SRCR DOMAIN.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF133845; AAD31850.1; -.
DR   EMBL; AF113248; AAF21966.1; -.
DR   HSSP; P01130; 1AJJ.
DR   MIM; 605236; -.
DR   InterPro; IPR000024; Fz_domain.
DR   InterPro; IPR001190; SRCR.
DR   InterPro; IPR001254; Trypsin.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR002172; LDL_recept_A.
DR   Pfam; PF01392; Fz; 2.
DR   Pfam; PF00057; ldl_recept_a; 7.
DR   Pfam; PF00089; trypsin; 1.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   PRINTS; PR00261; LDLRECEPTOR.
DR   SMART; SM00063; FRI; 2.
DR   SMART; SM00192; LDLa; 7.
DR   SMART; SM00202; SR; 1.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS50038; FZ; 2.
DR   PROSITE; PS01209; LDLRA_1; 6.
DR   PROSITE; PS50068; LDLRA_2; 7.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; FALSE_NEG.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
DR   PROSITE; PS00420; SRCR_1; FALSE_NEG.
DR   PROSITE; PS50287; SRCR_2; FALSE_NEG.
KW   Hydrolase; Serine protease; Transmembrane; Signal-anchor;
KW   Glycoprotein; Repeat.
FT   DOMAIN       1     45       CYTOPLASMIC (POTENTIAL).
FT   TRANSMEM     46    66       SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN)
FT                               (POTENTIAL).
FT   DOMAIN       67    1042     EXTRACELLULAR (POTENTIAL).
FT   DOMAIN       134   259      FZ 1.
FT   DOMAIN       268   304      LDL-RECEPTOR CLASS A 1.
FT   DOMAIN       305   340      LDL-RECEPTOR CLASS A 2.
FT   DOMAIN       341   377      LDL-RECEPTOR CLASS A 3.
FT   DOMAIN       378   415      LDL-RECEPTOR CLASS A 4.
FT   DOMAIN       450   573      FZ 2.
FT   DOMAIN       579   614      LDL-RECEPTOR CLASS A 5.
FT   DOMAIN       615   653      LDL-RECEPTOR CLASS A 6.
FT   DOMAIN       654   690      LDL-RECEPTOR CLASS A 7.
FT   DOMAIN       690   786      SRCR.
FT   DOMAIN       802   1042     SERINE PROTEASE.
FT   ACT_SITE     843   843      CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE     892   892      CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE     985   985      CHARGE RELAY SYSTEM.
FT   DISULFID     790   912      BY SIMILARITY.
FT   DISULFID     828   844      BY SIMILARITY.
FT   DISULFID     955   970      BY SIMILARITY.
FT   DISULFID     981   1010     BY SIMILARITY.
FT   CARBOHYD     80    80       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     104   104      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     135   135      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     141   141      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     231   231      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     245   245      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     251   251      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     305   305      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     320   320      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     376   376      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     413   413      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     446   446      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     451   451      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     469   469      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     567   567      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     651   651      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     697   697      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     761   761      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     1022  1022     N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   MUTAGEN      985   985      S->A: LOSS OF ACTIVITY.
FT   CONFLICT     854   854      W -> R (IN REF. 2).
FT   CONFLICT     876   876      K -> R (IN REF. 2).
SQ   SEQUENCE   1042 AA;  116564 MW;  7705398EBB607AD2 CRC64;
```

```
  Query Match             14.8%;  Score 692;  DB 1;  Length 1042;
  Best Local Similarity   26.8%;  Pred. No. 3.2e-39;
  Matches  206;  Conservative   95;  Mismatches  246;  Indels  222;  Gaps   31;

Qy    202 DSKTVQRTQDNSCSFGLHARGVELMRFTTPGFPDSPYPAHARCQWALRGDAD-------- 253
          |   | :: : :||   |::|:   |          |:  :|     || |
Db    365 DHDCVDKSDEVNCS--CHSQGLVECR------NGQCIPSTFQCD----GDEDCKDGSDEE 412

Qy    254 --SVLSLTFRSFD--------LASCDERGSDLVTVYNTL---SPMEPHALVQLCGTYP-- 298
            ||:  : :  |         | ||   || |    |:|    |  ||  | :||   |
Db    413 NCSVIQTSCQEGDQRCLYNPCLDSCG--GSSLCDPNNSLNNCSQCEPITL-ELCMNLPYN 469

Qy    299 ----PSY---------NLTFHSSQ-----------------NVLLITLITNTERRHPGF 327
              |:|         :::: ||                    :|:     ||  | |
Db    470 STSYPNYFGHRTQKEASISWESSLFPALVQTNCYKYLMFFSCTILVPKCDVNTGERIP-- 527

Qy    328 EATFFQLPRMSSCGGRLRKAQGTFNSPYYPGHYPPNIDCT-WNIEVPNNQHVKVSFKFFY 386
```

```
Db    528 -------PCRALCEHSKERCESVLG--IVGLQWPEDTDCSQFPEENSDNQ---------- 568

Qy    387 LLEPGVPAGTC--PKDYVEINGEKYCGERSQFVVTSNSNKITVRFHSDQSYTDTGFLAEY 444

Db    569 ---------TCLMPDEYVE---------------------------------------- 578

Qy    445 LSYDSSDPCPGQFTCRTGRCIRKELRCDGWADCTDHSDELNCSCDAGHQFTC-KNKFCKP 503

Db    579 ------ECSPSHFKCRSGQCVLISKRCDGQADCDDDSDEENCGCKERDLWECPSNKQCLK 632

Qy    504 LFWVCDSVNDCGDNSDEQGCS-CPAQTFRCSNGKCLSKSQQCNGKDDCGDGSDEASCPKV 562

Db    633 HTVICDGFPDCPDYMDEKNCSFCQDDELECANHACVSRDLWCDGEADCSDSSDEWDCVTL 692

Qy    563 NV------------------------------VTC------------------------T 568

Db    693 SINVNSSSFLMVHRAATEHHVCADGWQEILSQLACKQMGLGEPSVTKLIQEQEKEPRWLT 752

Qy    569 KHT-YRCLNGLCLSK---GNPECDGKEDCSDGSDEKDCDCGLRSFTR-QARVVGGTDADE 623

Db    753 LHSNWESLNGTTLHELLVNGQSCESRSKISLLCTKQ--DCGRRPAARMNKRILGGRTSRP 810

Qy    624 GEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYIDDRGFRYSDPTQWTAFLGLHDQSQ 683

Db    811 GRWPWQCSLQSEPSGHICGCVLIAKKWVLTVAHCF-EGR----ENAAVWKVVLGINNLDH 865

Qy    684 RSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKPAEYSSMVRPICLPDASHVFPAGK 743

Db    866 PSV-FMQTRFVKTIILHPRYSRAVVDYDISIVELSEDISETGYVRPVCLPNPEQWLEPDT 924

Qy    744 AIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLLPQQ-ITPRMMCVGFLSGGVDSCQ 802

Db    925 YCYITGWGH--MGNKMPFKLQEGEVRIISLEHCQSYFDMKTITTRMICAGYESGTVDSCM 982

Qy    803 GDSGGPLSSVEADGRIFQAGVVSWGDGCAQR-NKPGVYTRLPLFRDWIK 850

Db    983 GDSGGPLVCEKPGGRWTLFGLTSWGSVCFSKVLGPGVYSNVSYFVEWIK 1031
```

```
RESULT   6
ENTK_HUMAN
ID   ENTK_HUMAN      STANDARD;      PRT;  1019 AA.
AC   P98073;
DT   01-FEB-1996 (Rel. 33, Created)
DT   01-FEB-1996 (Rel. 33, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ENTEROPEPTIDASE PRECURSOR (EC 3.4.21.9) (ENTEROKINASE).
GN   PRSS7 OR ENTK.
OS   Homo sapiens (Human).
OC     karyota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC     mmalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Duodenum;
RX   MEDLINE=95234679; PubMed=7718557;
RA   Kitamoto Y., Veile R.A., Donis-Keller H., Sadler J.E.;
RT   "cDNA sequence and chromosomal localization of human enterokinase,
RT   the proteolytic activator of trypsinogen.";
RL   Biochemistry 34:4562-4568(1995).
RN   [2]
RP   SEQUENCE FROM N.A.
RA   Holzinger A., Buck C., Maier E.M., Mayerhofer P.U., Roscher A.A.,
RA   Sadler J.E., Hadorn H.B.;
RT   "Genomic organization of the human enteropeptidase.";
RL   Submitted (JUN-1999) to the EMBL/GenBank/DDBJ databases.
RN   [3]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=20289799; PubMed=10830953;
RA   Hattori M., Fujiyama A., Taylor T.D., Watanabe H., Yada T.,
RA   Park H.-S., Toyoda A., Ishii K., Totoki Y., Choi D.-K., Soeda E.,
RA   Ohki M., Takagi T., Sakaki Y., Taudien S., Blechschmidt K., Polley A.,
RA   Menzel U., Delabar J., Kumpf K., Lehmann R., Patterson D.,
RA   Reichwald K., Rump A., Schillhabel M., Schudy A., Zimmermann W.,
RA   Rosenthal A., Kudoh J., Shibuya K., Kawasaki K., Asakawa S.,
RA   Shintani A., Sasaki T., Nagamine K., Mitsuyama S., Antonarakis S.E.,
RA   Minoshima S., Shimizu N., Nordsiek G., Hornischer K., Brandt P.,
RA   Scharfe M., Schoen O., Desario A., Reichelt J., Kauer G., Bloecker H.,
RA   Ramser J., Beck A., Klages S., Hennig S., Riesselmann L., Dagand E.,
RA   Wehrmeyer S., Borzym K., Gardiner K., Nizetic D., Francis F.,
RA   Lehrach H., Reinhardt R., Yaspo M.-L.;
RT   "The DNA sequence of human chromosome 21.";
RL   Nature 405:311-319(2000).
RN   [4]
RP   SEQUENCE OF 749-1019 FROM N.A.
RC   TISSUE=Duodenum;
RX   MEDLINE=94329561; PubMed=8052624;
RA   Kitamoto Y., Yuan X., Wu Q., McCourt D.W., Sadler J.E.;
RT   "Enterokinase, the initiator of intestinal digestion, is a mosaic
RT   protease composed of a distinctive assortment of domains.";
RL   Proc. Natl. Acad. Sci. U.S.A. 91:7588-7592(1994).
CC   -!- FUNCTION: RESPONSIBLE FOR INITIATING ACTIVATION OF PANCREATIC
CC       PROTEOLYTIC PROENZYMES (TRYPSIN, CHYMOTRYPSIN AND CARBOXYPEPTIDASE
CC       A). IT CATALYZES THE CONVERSION OF TRYPSINOGEN TO TRYPSIN WHICH IN
CC       TURN ACTIVATES OTHER PROENZYMES INCLUDING CHYMOTRYPSINOGEN,
CC       PROCARBOXYPEPTIDASES, AND PROELASTASES.
CC   -!- CATALYTIC ACTIVITY: SELECTIVE CLEAVAGE OF 6-LYS-|-ILE-7 BOND IN
CC       TRYPSINOGEN.
CC   -!- SUBUNIT: HETERODIMER OF A CATALYTIC (LIGHT) CHAIN AND A
CC       MULTIDOMAIN (HEAVY) CHAIN LINKED BY A DISULFIDE BOND.
CC   -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN (PROBABLE).
CC   -!- TISSUE SPECIFICITY: INTESTINAL BRUSH BORDER.
CC   -!- PTM: THE CHAINS ARE DERIVED FROM A SINGLE PRECURSOR THAT IS
CC       CLEAVED BY A TRYPSIN-LIKE PROTEASE.
CC   -!- DISEASE: DEFECTS IN PRSS7 CAUSE LIFE-THREATENING INTESTINAL
CC       MALABSORPTION CHARACTERIZED BY DIARRHEA AND FAILURE TO THRIVE.
CC   -!- SIMILARITY: CONTAINS 2 LDL-RECEPTOR CLASS A DOMAINS.
CC   -!- SIMILARITY: CONTAINS 2 CUB DOMAINS.
CC   -!- SIMILARITY: CONTAINS 1 SEA DOMAIN.
CC   -!- SIMILARITY: CONTAINS 1 SRCR DOMAIN.
CC   -!- SIMILARITY: CONTAINS 1 MAM DOMAIN.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; U09860; AAC50138.1; -.
DR   EMBL; Y19124; CAB65555.1; -.
DR   EMBL; Y19125; CAB65555.1; JOINED.
DR   EMBL; Y19126; CAB65555.1; JOINED.
DR   EMBL; Y19127; CAB65555.1; JOINED.
DR   EMBL; Y19128; CAB65555.1; JOINED.
DR   EMBL; Y19129; CAB65555.1; JOINED.
DR   EMBL; Y19130; CAB65555.1; JOINED.
DR   EMBL; Y19131; CAB65555.1; JOINED.
DR   EMBL; Y19132; CAB65555.1; JOINED.
DR   EMBL; Y19133; CAB65555.1; JOINED.
DR   EMBL; Y19134; CAB65555.1; JOINED.
DR   EMBL; Y19135; CAB65555.1; JOINED.
DR   EMBL; Y19136; CAB65555.1; JOINED.
DR   EMBL; Y19137; CAB65555.1; JOINED.
DR   EMBL; Y19138; CAB65555.1; JOINED.
DR   EMBL; Y19139; CAB65555.1; JOINED.
DR   EMBL; Y19140; CAB65555.1; JOINED.
DR   EMBL; Y19141; CAB65555.1; JOINED.
DR   EMBL; Y19142; CAB65555.1; JOINED.
DR   EMBL; Y19143; CAB65555.1; JOINED.
DR   EMBL; AL163218; CAB90392.1; -.
DR   EMBL; AL163217; CAB90389.1; -.
```

```
DR   HSSP; P00763; 1DPO.
DR   MEROPS; S01.156; -.
DR   MIM; 226200; -.
DR   InterPro; IPR000859; CUB.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR002172; LDL_recept_A.
DR   InterPro; IPR000998; MAM.
DR   InterPro; IPR000082; SEA.
DR   InterPro; IPR001190; SRCR.
DR   InterPro; IPR001254; Trypsin.
DR   Pfam; PF00431; CUB; 2.
DR   Pfam; PF00057; ldl_recept_a; 2.
DR   Pfam; PF00629; MAM; 1.
DR   Pfam; PF01390; SEA; 1.
DR   Pfam; PF00530; SRCR; 1.
DR   Pfam; PF00089; trypsin; 1.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   SMART; SM00042; CUB; 2.
DR   SMART; SM00192; LDLa; 2.
DR   SMART; SM00137; MAM; 1.
DR   SMART; SM00200; SEA; 1.
DR   SMART; SM00202; SR; 1.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS01180; CUB; 2.
DR   PROSITE; PS01209; LDLRA_1; 2.
DR   PROSITE; PS50068; LDLRA_2; 2.
DR   PROSITE; PS00740; MAM_1; 1.
DR   PROSITE; PS50060; MAM_2; 1.
DR   PROSITE; PS50024; SEA; 1.
DR   PROSITE; PS00420; SRCR_1; FALSE_NEG.
DR   PROSITE; PS50287; SRCR_2; 1.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Signal-anchor; Glycoprotein; Myristate; Hydrolase;
KW   Serine protease; Zymogen; Transmembrane; Repeat.
FT   CHAIN         1    784       NON-CATALYTIC CHAIN (HEAVY CHAIN).
FT   CHAIN       785   1019       CATALYTIC CHAIN (LIGHT CHAIN).
FT   TRANSMEM     19     47       SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN).
FT   DOMAIN       52    169       SEA.
FT   DOMAIN      182    223       LDL-RECEPTOR CLASS A 1.
FT   DOMAIN      225    334       CUB 1.
FT   DOMAIN      342    504       MAM.
FT   DOMAIN      524    634       CUB 2.
FT   DOMAIN      641    679       LDL-RECEPTOR CLASS A 2.
FT   DOMAIN      678    771       SRCR.
FT   DOMAIN      785   1019       SERINE PROTEASE.
FT   ACT_SITE    825    825       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    876    876       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    971    971       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   LIPID         2      2       MYRISTATE (POTENTIAL).
FT   DISULFID    184    197       BY SIMILARITY.
FT   DISULFID    191    210       BY SIMILARITY.
FT   DISULFID    204    221       BY SIMILARITY.
FT   DISULFID    643    655       BY SIMILARITY.
FT   DISULFID    650    668       BY SIMILARITY.
FT   DISULFID    662    677       BY SIMILARITY.
FT   DISULFID    772    896       INTERCHAIN (BY SIMILARITY).
FT   DISULFID    810    826       BY SIMILARITY.
FT   DISULFID    910    977       BY SIMILARITY.
FT   DISULFID    941    956       BY SIMILARITY.
FT   DISULFID    967    995       BY SIMILARITY.
FT   CARBOHYD    116    116       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    147    147       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    179    179       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    328    328       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    335    335       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    388    388       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    440    440       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    470    470       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    503    503       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    534    534       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    630    630       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    682    682       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    706    706       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    725    725       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    848    848       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    887    887       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    909    909       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    949    949       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT    134    134       Q -> E (IN REF. 3).
FT   CONFLICT    732    732       S -> P (IN REF. 3).
FT   CONFLICT    754    771       SQQCLQDSLIRLQCNHKS -> RRNAKNEIDALSPIILIA
FT                                (IN REF. 3).
SQ   SEQUENCE   1019 AA;  112923 MW;  B6AAA245F6D4A563 CRC64;


  Query Match             14.6%;  Score 682.5;  DB 1;  Length 1019;
  Best Local Similarity   32.5%;  Pred. No. 1.4e-38;
  Matches  250;  Conservative   80;  Mismatches  202;  Indels   91;  Gaps   20;

Qy     333 QLPRMSSCGG--RLRKAQGTFNSPYYPGHYPPNIDCTWNIEVPNNQHVKVSFKFFYLLEP 390
           :||  : |||   | :  ||:|  :|  ||    | | :     ::::: |: | |
Db     519 ELP--TDCGGPFELWEPNTTFSSTNFPNSYPNLAFCVWILNAQKGKNIQLHFQEFDLEN- 575

Qy     391 GVPAGTCPKDYVEI-NGEK--------YCGERSQFVVTSNSNKITVRFHSDQSYTDTGFL 441
             | ||| :||:         | |     | | :|::||    ::       ||
Db     576 -------INDVVEIRDGEEADSLLLAVYTGPGPVKDVFSTTNRMTVLLITNDVLARGGFK 628

Qy     442 AEYLS---YDSSDPCPG-QFTCRTGRCIRKELRCDGWADCTDHSDELNCSCDAGHQF--- 494
           | : :        :||    | |: | |:     |||   | | ||| :|      :|
Db     629 ANFTTGYHLGIPEPCKADHFQCKNGECVPLVNLCDGHLHCEDGSDEADCV-----RFFNG 683

Qy     495 TCKN----KFCKPLFWVCDSVNDCGDNSDEQGCSCPAQTFRCSNGKCLSKSQQCNGKDDC 550
           |  |    :|     |        | :|           | : ||  |
Db     684 TTNNNGLVRFRIQSIW----HTACAEN----------WTTQISNDVC----------QLL 719

Qy     551 GDGSDEASCPKVNVVTCTKHTYRCLNGLCLSKGNPECDG------KEDCSDGS------D 598
           | ||  :| |  :           :|    | |   ||       : |   |      :
Db     720 GLGSGNSSKPIFST----------DGGPFVKLNTAPDGHLILTPSQQCLQDSLIRLQCN 768

Qy     599 EKDCDCGLRSFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCY 658
           | |   | :     ::|||::| || ||| | |:  |   :||||||:| :|||||||
Db     769 HKSCGKKLAAQDITPKIVGGSNAKEGAWPWVVGLY-YGGRLLCGASLVSSDWLVSAAHCV 827

Qy     659 IDDRGFRYSDPTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELE 718
                |  :|::||| |||| :|  ::|    | :  |: :| :|    | |||:: ||
Db     828 YG----RNLEPSKWTAILGLHMKSNLTSPQTVPRLIDEIVINPHYNRRRKDNDIAMMHLE 883

Qy     719 KPAEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCEN 778
              |:  ::|||||: : ||| |:   : |||   | || | |||: :: :::   |:
Db     884 FKVNYTDYIQPICLPEENQVFPPGRNCSIAGWGTVVYQGTTANILQEADVPLLSNERCQQ 943

Qy     779 LLPQ-QITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPG 837
           :|:  ||   |:| |:  ||:||||||||||   | : | | ||| |:|  ||   |:||
Db     944 QMPEYNITENMICAGYEEGGIDSCQGDSGGPLMCQE-NNRWFLAGVTSFGYKCALPNRPG 1002

Qy     838 VYTRLPLFRDWIK 850
           || |:  | :||:
Db    1003 VYARVSRFTEWIQ 1015


RESULT   7
ENTK_MOUSE
ID   ENTK_MOUSE     STANDARD;      PRT;  1069 AA.
AC   P97435;
DT   01-NOV-1997 (Rel. 35, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ENTEROPEPTIDASE (EC 3.4.21.9) (ENTEROKINASE).
GN   PRSS7 OR ENTK.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
```

```
RP   SEQUENCE FROM N.A.
RC   STRAIN=C57BL/6; TISSUE=Duodenum;
RX   MEDLINE=98147142; PubMed=9486188;
RA   Yuan X., Zheng X., Lu D., Rubin D.C., Pung C.Y.M., Sadler J.E.;
RT   "Structure of murine enterokinase (enteropeptidase) and expression in
RT   small intestine during development.";
RL   Am. J. Physiol. 274:G342-G349(1998).
CC   -!- FUNCTION: RESPONSIBLE FOR INITIATING ACTIVATION OF PANCREATIC
CC       PROTEOLYTIC PROENZYMES (TRYPSIN, CHYMOTRYPSIN AND CARBOXYPEPTIDASE
CC       A). IT CATALYZES THE CONVERSION OF TRYPSINOGEN TO TRYPSIN WHICH IN
CC       TURN ACTIVATES OTHER PROENZYMES INCLUDING CHYMOTRYPSINOGEN,
CC       PROCARBOXYPEPTIDASES, AND PROELASTASES (BY SIMILARITY).
CC   -!- CATALYTIC ACTIVITY: SELECTIVE CLEAVAGE OF 6-LYS-|-ILE-7 BOND IN
CC       TRYPSINOGEN.
CC   -!- SUBUNIT: HETERODIMER OF A CATALYTIC (LIGHT) CHAIN AND A
CC       MULTIDOMAIN (HEAVY) CHAIN LINKED BY A DISULFIDE BOND (BY
CC       SIMILARITY).
CC   -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN (PROBABLE).
CC   -!- PTM: THE CHAINS ARE DERIVED FROM A SINGLE PRECURSOR THAT IS
CC       CLEAVED BY A TRYPSIN-LIKE PROTEASE (BY SIMILARITY).
CC   -!- SIMILARITY: CONTAINS 2 LDL-RECEPTOR CLASS A DOMAINS.
CC   -!- SIMILARITY: CONTAINS 2 CUB DOMAINS.
CC     - SIMILARITY: CONTAINS 1 SEA DOMAIN.
CC     - SIMILARITY: CONTAINS 1 SRCR DOMAIN.
CC   -!- SIMILARITY: CONTAINS 1 MAM DOMAIN.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; U73378; AAB37317.1; -.
DR   HSSP; P00763; 1DPO.
DR   MEROPS; S01.156; -.
DR   MGD; MGI:1197523; Prss7.
DR   InterPro; IPR000859; CUB.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR002172; LDL_recept_A.
DR   InterPro; IPR000998; MAM.
DR   InterPro; IPR000082; SEA.
DR   InterPro; IPR001190; SRCR.
DR   InterPro; IPR001254; Trypsin.
DR   Pfam; PF00431; CUB; 2.
DR   Pfam; PF00057; ldl_recept_a; 2.
DR   Pfam; PF00629; MAM; 1.
DR   Pfam; PF01390; SEA; 1.
DR   Pfam; PF00530; SRCR; 1.
DR   Pfam; PF00089; trypsin; 1.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   SMART; SM00042; CUB; 2.
DR   SMART; SM00192; LDLa; 2.
DR   SMART; SM00137; MAM; 1.
DR   SMART; SM00200; SEA; 1.
DR   SMART; SM00202; SR; 1.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS01180; CUB; 2.
DR   PROSITE; PS01209; LDLRA_1; 2.
DR   PROSITE; PS50068; LDLRA_2; 2.
DR   PROSITE; PS00740; MAM_1; 1.
DR   PROSITE; PS50060; MAM_2; 1.
DR   PROSITE; PS50024; SEA; 1.
DR   PROSITE; PS00420; SRCR_1; FALSE_NEG.
DR   PROSITE; PS50287; SRCR_2; 1.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Signal-anchor; Glycoprotein; Myristate; Hydrolase;
KW   Serine protease; Zymogen; Transmembrane; Repeat.
```

```
FT   CHAIN         1    829       NON-CATALYTIC CHAIN (HEAVY CHAIN).
FT   CHAIN       830   1069       CATALYTIC CHAIN (LIGHT CHAIN).
FT   TRANSMEM     19     47       SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN).
FT   DOMAIN       52    169       SEA.
FT   DOMAIN      227    268       LDL-RECEPTOR CLASS A 1.
FT   DOMAIN      270    379       CUB 1.
FT   DOMAIN      387    549       MAM.
FT   DOMAIN      569    679       CUB 2.
FT   DOMAIN      686    724       LDL-RECEPTOR CLASS A 2.
FT   DOMAIN      723    816       SRCR.
FT   DOMAIN      830   1069       SERINE PROTEASE.
FT   ACT_SITE    874    874       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    925    925       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE   1021   1021       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   LIPID         2      2       MYRISTATE (POTENTIAL).
FT   DISULFID    229    242       BY SIMILARITY.
FT   DISULFID    236    255       BY SIMILARITY.
FT   DISULFID    249    266       BY SIMILARITY.
FT   DISULFID    688    700       BY SIMILARITY.
FT   DISULFID    695    713       BY SIMILARITY.
FT   DISULFID    707    722       BY SIMILARITY.
FT   DISULFID    817    945       INTERCHAIN (BY SIMILARITY).
FT   DISULFID    859    875       BY SIMILARITY.
FT   DISULFID    959   1027       BY SIMILARITY.
FT   DISULFID    991   1006       BY SIMILARITY.
FT   DISULFID   1017   1045       BY SIMILARITY.
FT   CARBOHYD    147    147       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    197    197       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    212    212       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    373    373       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    380    380       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    433    433       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    515    515       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    579    579       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    675    675       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    727    727       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    751    751       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    770    770       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    791    791       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    897    897       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    936    936       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    999    999       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1069 AA;  118735 MW;  E62549E463743C3D CRC64;
```

```
  Query Match           14.5%;  Score 676.5;  DB 1;  Length 1069;
  Best Local Similarity 29.2%;  Pred. No. 3.7e-38;
  Matches  186;  Conservative   93;  Mismatches  220;  Indels  137;  Gaps   23;

Qy    258 LTFRSFDLASCDERGSDLVTVYNTLSPMEPHALVQLCGTYPPSYNLTFHSSQNVLLITLI 317
          : | :|    |     | ::: | :    |:    |  | ||
Db    522 VVFNAFRNRGCSTIALDDISLTNGICSQSPYPEPTLVPTPPP----------------- 563

Qy    318 TNTERRHPGFEATFFQLPRMSSCGG--RLRKAQGTFNSPYYPGHYPPNIDCTWNIEVPNN 375
                    :||  : |||   | :   ||:|| :|  ||    | ||:
Db    564 --------------ELP--TDCGGPFELWEPNSTFSSPNFPDKYPNQASCIWNLNAQRG 606

Qy    376 QHVKVSFKFFYLLEPGVPAGTCPKDYVEI--NGE-------KYCGERSQFVVTSNSNKIT 426
          ::::: |: | |            | ||:   ||        | |     : | :|::|
Db    607 KNIQLHFQEFDLEN--------INDVVEVRDGGEFDSLLLAVYTGPGPVKDLFSTTNRMT 658

Qy    427 VRFHSDQSYTDTGFLAEYLS---YDSSDPC-PGQFTCRTGRCIRKELRCDGWADCTDHSD 482
          | | ::      || | : | ·    :||   :| |: | ||      || :  | | ||
Db    659 VIFTTNMETRRKGFKANFTSGYYLGIPEPCQDDEFQCKDGNCIPLGNLCDSYPHCRDGSD 718

Qy    483 ELNC--------SCDAGHQFTCKNKFCKPLFWVCDSVNDCGDNSDEQGCSCPAQTFRCSN 534
          | :|        | :   ||   :   |       | :|           | : ||
Db    719 EASCVRFLNGTRSNNGLVQFNIHS------IWHI----ACAEN----------WTTQISN 758

Qy    535 GKCLSKSQQCNGKDDCGDGSDEASCP----------KVNVVTCTKHTYRCLNGLCLSKGN 584
            |              | ||  :| |          :||         :  ||  :   :
Db    759 EVC----------HLLGLGSANSSMPISSTGGGPFVRVN---------QAPNGSLILTPS 799
```

```
Qy    585 PECDGKE----DCSDGSDEKDCDCGLRSFTRQA--RVVGGTDADEGEWPWQVSLHALGQG 638
          :|         |: |     .|| : |::   ::|||:||  | ||| |:|:   :
Db    800 LQCSQDSLILLQCNHKS------CGEKKVTQKVSPKIVGGSDAQAGAWPWVVALYHRDRS 853

Qy    639 H---ICGASLISPNWLVSAAHCYIDDRGFRYSDPTQWTAFLGLHDQSQRSAPGVQERRLK 695
          :|||||:| :|||||||        |  |||:||| |||| ||  ::| |  | :
Db    854 TDRLLCGASLVSSDWLVSAAHCVYR----RNLDPTRWTAVLGLHMQSNLTSPQVVRRVVD 909

Qy    696 RIISHPFFNDFTFDYDIALLELEKPAEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQY 755
          :|: :| ::      |||:: ||    |:  ::|||||: : :|  |:   : |||: :
Db    910 QIVINPHYDRRRKVNDIAMMHLEFKVNYTDYIQPICLPEENQIFIPGRTCSIAGWGYDKI 969

Qy    756 -GGTGALILQKGEIRVINQTTCENLLPQ-QITPRMMCVGFLSGGVDSCQGDSGGPLSSVE 813
           |:   :|:: :: :|:   |:  ||:  ||  |:| |:  ||:|||||||||||   |
Db    970 NAGSTVDVLKEADVPLISNEKCQQQLPEYNITESMICAGYEEGGIDSCQGDSGGPLMCQE 1029

Qy    814 ADGRIFQAGVVSWGDGCAQRNKPGVYTRLPLFRDWI 849
          : | |  || |:|  ||  | |||| |:  | :||
Db   1030 -NNRWFLVGVTSFGVQCALPNHPGVYVRVSQFIEWI 1064
```

```
RESULT   8
CORI_MOUSE
ID   CORI_MOUSE     STANDARD;      PRT;  1113 AA.
AC   Q9Z319;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   ATRIAL NATRIUTERIC PEPTIDE-CONVERTING ENZYME (EC 3.4.21.-) (PRO-ANP-
DE   CONVERTING ENZYME) (CORIN) (LOW DENSITY LIPOPROTEIN RECEPTOR RELATED
DE   PROTEIN 4).
GN   CRN OR LRP4.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=98429596; PubMed=9756624;
RA   Tomita Y., Kim D.-H., Magoori K., Fujino T., Yamamoto T.T.;
RT   "A novel low-density lipoprotein receptor-related protein with type II
RT   membrane protein-like structure is abundant in heart.";
RL   J. Biochem. 124:784-789(1998).
CC   -!- FUNCTION: CONVERTS PRO-ANP TO ANP. CLEAVES PRO-ANP SPECIFICALLY
CC       BETWEEN ARG-122 AND SER-123 (BY SIMILARITY).
CC   -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN.
CC   -!- TISSUE SPECIFICITY: HIGHLY EXPRESSED IN HEART.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   -!- SIMILARITY: CONTAINS 7 LDL-RECEPTOR CLASS A DOMAINS.
CC   -!- SIMILARITY: CONTAINS 2 FRIZZLED (FZ) DOMAINS.
CC   -!- SIMILARITY: CONTAINS 1 SRCR DOMAIN.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AB013874; BAA34371.1; -.
DR   HSSP; P01130; 1AJJ.
DR   MGD; MGI:1349451; Lrp4.
DR   InterPro; IPR000024; Fz_domain.
DR   InterPro; IPR001190; SRCR.
DR   InterPro; IPR001254; Trypsin.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR002172; LDL_recept_A.
DR   Pfam; PF01392; Fz; 2.
DR   Pfam; PF00057; ldl_recept_a; 7.
DR   Pfam; PF00089; trypsin; 1.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   PRINTS; PR00261; LDLRECEPTOR.
DR   SMART; SM00063; FRI; 2.
DR   SMART; SM00192; LDLa; 7.
DR   SMART; SM00202; SR; 1.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS50038; FZ; 2.
DR   PROSITE; PS01209; LDLRA_1; 6.
DR   PROSITE; PS50068; LDLRA_2; 7.
DR   PROSITE; PS00420; SRCR_1; FALSE_NEG.
DR   PROSITE; PS50287; SRCR_2; 1.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; FALSE_NEG.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Hydrolase; Serine protease; Transmembrane; Signal-anchor;
KW   Glycoprotein; Repeat.
FT   DOMAIN        1    112       CYTOPLASMIC (POTENTIAL).
FT   TRANSMEM    113    133       SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN)
FT                                (POTENTIAL).
FT   DOMAIN      134   1113       EXTRACELLULAR (POTENTIAL).
FT   DOMAIN      201    327       FZ 1.
FT   DOMAIN      336    372       LDL-RECEPTOR CLASS A 1.
FT   DOMAIN      373    408       LDL-RECEPTOR CLASS A 2.
FT   DOMAIN      409    445       LDL-RECEPTOR CLASS A 3.
FT   DOMAIN      446    483       LDL-RECEPTOR CLASS A 4.
FT   DOMAIN      518    641       FZ 2.
FT   DOMAIN      647    682       LDL-RECEPTOR CLASS A 5.
FT   DOMAIN      683    721       LDL-RECEPTOR CLASS A 6.
FT   DOMAIN      722    757       LDL-RECEPTOR CLASS A 7.
FT   DOMAIN      758    853       SRCR.
FT   DOMAIN      869   1113       SERINE PROTEASE.
FT   ACT_SITE    910    910       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    959    959       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE   1052   1052       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   DISULFID    857    979       BY SIMILARITY.
FT   DISULFID    895    911       BY SIMILARITY.
FT   DISULFID   1022   1037       BY SIMILARITY.
FT   DISULFID   1048   1077       BY SIMILARITY.
FT   CARBOHYD    147    147       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    202    202       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    208    208       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    298    298       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    317    317       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    373    373       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    411    411       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    444    444       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    481    481       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    519    519       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    537    537       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    635    635       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    719    719       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    765    765       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    828    828       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    970    970       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   1089   1089       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   1113 AA;  122984 MW;  B845B2C5F20DD8EC CRC64;
```

```
  Query Match           14.2%;  Score 663.5;  DB 1;  Length 1113;
  Best Local Similarity  32.7%;  Pred. No. 2.9e-37;
  Matches  158;  Conservative   67;  Mismatches  187;  Indels   71;  Gaps   16;

Qy    428 RFHSDQSYTDTGFLAEYLSYDSSDPCPGQFTCRTGRCIRKELRCDGWADCTDHSDELNCS 487
          :|  : |   |  |    : |  :  |  | ||:|||:    |||| ||| | ||| ||
Db    627 QFPEESSDNQTCLLP---NEDVEECSPSHFKCRSGRCVLGSRRCDGQADCDDDSDEENCG 683

Qy    488 CDAGHQFTCK-NKFCKPLFWVCDSVNDCGDNSDEQGCS-CPAQTFRCSNGKCLSKSQQCN 545
          |    : |  || |    :||   || |: ||: || |     |:| :|: : . |:
Db    684 CKERALWECPFNKQCLKHTLICDGFPDCPDSMDEKNCSFCQDNELECANHECVPRDLWCD 743

Qy    546 GKDDCGDGSDEASCPKVN----------VVTCTKHTYRCLNG--------LC-------- 579
          |  || | |||  |  :: .          |    |  : | :|        |
Db    744 GWVDCSDSSDEWGCVTLSKNGNSSSLLTVHKSAKEHHVCADGWRETLSQLACKQMGLGEP 803
```

```
Qy     580 -LSKGNPECDGKE--------DCSDGSDEKD-----------------C---DCGLRSFT 610
           ::|  |  :|::       :  :||  ::                   |   ||| |
Db     804 SVTKLIPGQEGQQWLRLYPNWENLNGSTLQELLVYRHSCPSRSEISLLCSKQDCGRRPAA 863

Qy     611 R-QARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYIDDRGFRYSDP 669
           |  |::||  :  | |||| || :   |||||  ||:  |::: |||: : |      |
Db     864 RMNKRILGGRTSRPGRWPWQCSLQSEPSGHICGCVLIAKKWVLTVAHCF-EGR----EDA 918

Qy     670 TQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEKPAEYSSMVRP 729
           |    |:::    |   :|. | :| |: || ::    ||||:::||      :| |||
Db     919 DVWKVVFGINNLDHPSG-FMQTRFVKTILLHPRYSRAVVDYDISVVELSDDINETSYVRP 977

Qy     730 ICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENLLPQQ-ITPRM 788
           :|||           ::|||||   |      ||:||:|:|    |::    : || ||
Db     978 VCLPSPEEYLEPDTYGYITGWGH--MGNKMPFKLQEGEVRIIPLEQCQSYFDMKTITNRM 1035

Qy     789 MCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQR-NKPGVYTRLPLFRD 847
           :| |: || |||| ||||||      |:    |: |||  |  :   ||||: :  |
Db    1036 ICAGYESGTVDSCMGDSGGPLVCERPGGQWTLFGLTSWGSVCFSKVLGPGVYSNVSYFVG 1095

Qy     848 WIK 850
           |||
Db     096 WIE 1098
```

```
RESULT   9
TMS2_MOUSE
ID   TMS2_MOUSE     STANDARD;      PRT;   490 AA.
AC   Q9JIQ8; Q9JKC4; Q9QY82;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   TRANSMEMBRANE PROTEASE, SERINE 2 (EC 3.4.21.-) (EPITHELIASIN) (PLASMIC
DE   TRANSMEMBRANE PROTEIN X).
GN   TMPRSS2.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A., AND TISSUE SPECIFICITY.
RC   STRAIN=BALB/C;
RX   MEDLINE=21104370; PubMed=11169526;
RA   Vaarala M.H., Porvari K.S., Kellokumpu S., Kyllonen A.P., Vihko P.T.;
RT   "Expression of transmembrane serine protease TMPRSS2 in mouse and
RT   human tissues.";
RL   J. Pathol. 193:134-140(2001).
RN   [2]
RP   SEQUENCE FROM N.A.
RA   Han J., Kim S.;
RT   "Putative transmembrane protease X.";
RL   Submitted (MAR-2000) to the EMBL/GenBank/DDBJ databases.
RN   [3]
RP   SEQUENCE FROM N.A.
RC   STRAIN=BALB/C;
RA   Jacquinet E.J., Rao N.V., Rao G.N., Hoidal J.R.;
RT   "A novel mosaic serine protease, epitheliasin.";
RL   Submitted (DEC-1998) to the EMBL/GenBank/DDBJ databases.
CC   -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN.
CC   -!- TISSUE SPECIFICITY: EXPRESSED MAINLY IN PROSTATE AND KIDNEY.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   -!- SIMILARITY: CONTAINS 1 SRCR DOMAIN.
CC   -!- SIMILARITY: CONTAINS 1 LDL-RECEPTOR CLASS A DOMAIN.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
```

```
DR   EMBL; AF199362; AAF97867.1; -.
DR   EMBL; AF243500; AAF64186.1; -.
DR   EMBL; AF113596; AAF21308.1; -.
DR   HSSP; P00761; 1AKS.
DR   MGD; MGI:1354381; Tmprss2.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR002172; LDL_recept_A.
DR   InterPro; IPR001190; SRCR.
DR   InterPro; IPR001254; Trypsin.
DR   Pfam; PF00057; ldl_recept_a; 1.
DR   Pfam; PF00089; trypsin; 1.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   SMART; SM00192; LDLa; 1.
DR   SMART; SM00202; SR; 1.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS01209; LDLRA_1; 1.
DR   PROSITE; PS50068; LDLRA_2; 1.
DR   PROSITE; PS00420; SRCR_1; FALSE_NEG.
DR   PROSITE; PS50287; SRCR_2; 1.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Hydrolase; Serine protease; Transmembrane; Signal-anchor.
FT   DOMAIN       1     83       CYTOPLASMIC (POTENTIAL).
FT   TRANSMEM     84    104      SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN)
FT                               (POTENTIAL).
FT   DOMAIN       105   490      EXTRACELLULAR (POTENTIAL).
FT   DOMAIN       111   149      LDL-RECEPTOR CLASS A.
FT   DOMAIN       150   242      SRCR.
FT   DOMAIN       254   490      SERINE PROTEASE.
FT   ACT_SITE     294   294      CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE     343   343      CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE     439   439      CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   SITE         253   254      CLEAVAGE (POTENTIAL).
FT   DISULFID      76   125      BY SIMILARITY.
FT   DISULFID     119   138      BY SIMILARITY.
FT   DISULFID     132   147      BY SIMILARITY.
FT   DISULFID     171   230      BY SIMILARITY.
FT   DISULFID     184   240      BY SIMILARITY.
FT   DISULFID     243   363      BY SIMILARITY.
FT   DISULFID     279   295      BY SIMILARITY.
FT   DISULFID     408   424      BY SIMILARITY.
FT   DISULFID     435   463      BY SIMILARITY.
FT   CARBOHYD     111   111      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     212   212      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD     474   474      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT     122   122      S -> L (IN REF. 3).
FT   CONFLICT     178   178      S -> N (IN REF. 3).
FT   CONFLICT     320   320      Y -> H (IN REF. 1).
FT   CONFLICT     474   474      N -> D (IN REF. 1).
SQ   SEQUENCE   490 AA;  53479 MW;  07D2B03EA4D8A1A9 CRC64;
```

```
  Query Match             12.8%;  Score 600;  DB 1;  Length 490;
  Best Local Similarity   34.3%;  Pred. No. 2.2e-33;
  Matches  137;  Conservative   62;  Mismatches  152;  Indels   48;  Gaps   11;

Qy     461 TGRCIRKELRCDGWADCTDHSDELNCSCDAGHQFTCKNKFCKPLFW--VCDSVNDCGDNS 518
           :| ||   | ||| | | :  ||  |    |  |  :    :   |  ||   :|  ::
Db     122 SGTCISSSLWCDGVAHCPNGEDENRCVRLYGQSFILQVYSSQRKAWYPVCQ--DDWSESY 179

Qy     519 DEQGCSCPAQTFRCSNGKCLSKSQQCNGKDDCGDGSDEASCPKVNVVTCTKHTYRCLNGL 578
                 |              | :           | |   |  |:||
Db     180 GRAAC------------KDMGYKNNFYSSQGIPDQSGATSFMKLNV------------- 213

Qy     579 CLSKGNPECDGKEDCSDGSDEK-----DC-DCGLRSFTRQARVVGGTDADEGEWPWQVSL 632
               | ||  :   |   ||    :       | :||:||  ||:|:||| :|  |:||||||
Db     214 --SSGNVDLYKKLYHSDSCSSRMVVSLRCIECGVRSVKRQSRIVGGLNASPGDWPWQVSL 271

Qy     633 HALGQGHICGASLISPNWLVSAAHCYIDDRGFRYSDPTQWTAFLGLHDQSQRSAPGVQER 692
           |  |  |:|| |:|:| |:|:|||||  :      | |  |||| |:  ||
Db     272 HVQGV-HVCGGSIITPEWIVTAAHCVEEP----LSGPRYWTAFAGILRQSLMFYG--SRH 324
```

```
Qy    693 RLKRIISHPFFNDFTFDYDIALLELEKPAEYSSMVRPICLPDASHVFPAGKAIWVTGWGH 752
          :::::|||| :: . | : |||||:|: | ::: :|:|:|||:   :    :  |::|||
Db    325 QVEKVISHPNYDSKTKNNDIALMKLQTPLAFNDLVKPVCLPNPGMMLDLDQECWISGWGA 384

Qy    753 TQYGGTGALILQKGEIRVINQTTCEN--LLPQQITPRMMCVGFLSGGVDSCQGDSGGPLS 810
          |   |  : :|    : :|  :.|:.:    ||| |:| ||| | ||||||||||||
Db    385 TYEKGKTSDVLNAAMVPLIERSKCNSKYIYNNLITPAMICAGFLQGSVDSCQGDSGGPLV 444

Qy    811 SVEADGRIFQAGVVSWGDGCAQRNKPGVYTRLPLFRDWI 849
          ::: :|  :  |  ||| |||:  :|||| : :| |||
Db    445 TLK-NGIWWLIGDTSWGSGCAKALRPGVYGNVTVFTDWI 482


RESULT 10
CRAR_MOUSE
ID   CRAR_MOUSE     STANDARD;      PRT;   704 AA.
AC   P98064;
DT   01-FEB-1996 (Rel. 33, Created)
DT   01-FEB-1996 (Rel. 33, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   COMPLEMENT-ACTIVATING COMPONENT OF RA-REACTIVE FACTOR PRECURSOR
DE   (EC 3.4.21.-) (RA-REACTIVE FACTOR SERINE PROTEASE P100) (RARF)
DE   (MANNAN-BINDING LECTIN SERINE PROTEASE 1).
GN   MASP1 OR CRARF.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.
RC   STRAIN=BALB/C; TISSUE=Liver;
RX   MEDLINE=94179811; PubMed=8133044;
RA   Takayama Y., Takada F., Takahashi A., Kawakami M.;
RT   "A 100-kDa protein in the C4-activating component of Ra-reactive
RT   factor is a new serine protease having module organization similar to
RT   C1r and C1s.";
RL   J. Immunol. 152:2308-2316(1994).
RN   [2]
RP   SEQUENCE OF 465-704 FROM N.A., AND PARTIAL SEQUENCE.
RC   STRAIN=BALB/C; TISSUE=Liver;
RX   MEDLINE=93176166; PubMed=8439319;
RA   Takahashi A., Takayama Y., Hatsuse H., Kawakami M.;
RT   "Presence of a serine protease in the complement-activating component
RT   of the complement-dependent bactericidal factor, RaRF, in mouse
RT   serum.";
RL   Biochem. Biophys. Res. Commun. 190:681-687(1993).
CC   -!- FUNCTION: COMPONENT OF THE BACTERICIDAL RA-REACTIVE FACTOR RARF
CC       WHICH SPECIFICALLY BINDS TO RA AND R2 POLYSACCHARIDES EXPRESSED BY
CC       CERTAIN ENTEROBACTERIA. IT TRIGGERS THE ACTIVATION OF COMPLEMENT
CC       CASCADE BY ACTIVATING THE C4 AND C2 COMPONENTS. IT ACTIVATES THE
CC       C4 COMPONENT BY CLEAVING THE ALPHA-CHAIN OF C4.
CC   -!- SUBUNIT: RARF CONSISTS OF A COMPLEMENT-ACTIVATING COMPONENT
CC       (CRARF) AND A POLYSACCHARIDE-BINDING (MANNOSE-BINDING) COMPONENT.
CC       CRARF IS AN HETERODIMER OF A HEAVY (P70) AND A LIGH CHAIN (29)
CC       LINKED BY A DISULFIDE BOND.
CC   -!- TISSUE SPECIFICITY: LIVER.
CC   -!- DOMAIN: CRARF HAS A MODULE ORGANIZATION SIMILAR TO C1R AND C1S.
CC   -!- SIMILARITY: CONTAINS 2 SUSHI (SCR) DOMAINS.
CC   -!- SIMILARITY: CONTAINS 2 CUB DOMAINS.
CC   -!- SIMILARITY: CONTAINS 1 EGF-LIKE DOMAIN.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; D16492; BAA03944.1; -.
DR   HSSP; P00736; 1APQ.
DR   MEROPS; S01.198; -.
DR   MGD; MGI:88492; Maspl.
DR   InterPro; IPR000152; Asx_hydroxyl.
DR   InterPro; IPR000859; CUB.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR001881; EGF_Ca.
DR   InterPro; IPR000436; Sushi_SCR_CCP.
DR   InterPro; IPR001254; Trypsin.
DR   Pfam; PF00431; CUB; 2.
DR   Pfam; PF00084; sushi; 2.
DR   Pfam; PF00089; trypsin; 1.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   SMART; SM00032; CCP; 2.
DR   SMART; SM00042; CUB; 2.
DR   SMART; SM00179; EGF_CA; 1.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS00010; ASX_HYDROXYL; 1.
DR   PROSITE; PS01180; CUB; 2.
DR   PROSITE; PS01186; EGF_2; 1.
DR   PROSITE; PS01187; EGF_CA; 1.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Hydrolase; Complement pathway; Serine protease; Protease;
KW   Glycoprotein; Sushi; Repeat; Signal; EGF-like domain; Hydroxylation.
FT   SIGNAL       1     24
FT   CHAIN       25    704       COMPLEMENT-ACTIVATING COMPONENT OF
FT                                 RA-REACTIVE FACTOR.
FT   CHAIN       25    453       70 KDA CHAIN OF P100 (P70).
FT   CHAIN      454    704       29 KDA CHAIN OF P100 (P29).
FT   DOMAIN      25    143       CUB 1.
FT   DOMAIN     144    187       EGF-LIKE, CALCIUM-BINDING (POTENTIAL).
FT   DOMAIN     190    302       CUB 2.
FT   DOMAIN     305    368       SUSHI 1.
FT   DOMAIN     371    438       SUSHI 2.
FT   DOMAIN     454    704       SERINE PROTEASE.
FT   ACT_SITE   495    495       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE   557    557       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE   651    651       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   MOD_RES    164    164       HYDROXYLATION (POTENTIAL).
FT   DISULFID    78     96       POTENTIAL.
FT   DISULFID   148    162       POTENTIAL.
FT   DISULFID   158    171       POTENTIAL.
FT   DISULFID   173    186       POTENTIAL.
FT   DISULFID   190    217       POTENTIAL.
FT   DISULFID   247    265       POTENTIAL.
FT   DISULFID   306    354       POTENTIAL.
FT   DISULFID   334    367       POTENTIAL.
FT   DISULFID   372    419       POTENTIAL.
FT   DISULFID   402    437       POTENTIAL.
FT   DISULFID   441    577       INTERCHAIN (POTENTIAL).
FT   DISULFID   619    636       POTENTIAL.
FT   DISULFID   647    677       POTENTIAL.
FT   CARBOHYD    54     54       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   183    183       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   390    390       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD   412    412       N-LINKED (GLCNAC. . .) (POTENTIAL).
SQ   SEQUENCE   704 AA;  79895 MW;  71F44F3012D2C67F CRC64;


  Query Match             12.5%;  Score 586.5;  DB 1;  Length 704;
  Best Local Similarity   25.3%;  Pred. No. 2.7e-32;
  Matches  191;  Conservative  107;  Mismatches  261;  Indels  197;  Gaps   31;

Qy    220 ARGVELM----RFTTPGFPDSPYPAHARCQWALRGDADSVLSLTFRSFDLASCDERGSDL 275
          |  |||     :  :||:||| ||: :   | :       : | |  |:| |      |
Db     24 AHTVELNEMFGQIQSPGYPDS-YPSDSEVTWNITVPEGFRIKLYFMHFNLESSYLCEYDY 82

Qy    276 VTVYNTLSPMEPHALVQLCGTYPPSYNLTFHSSQNVLL-------ITLIT--NTERRHPG 326
          | |       |   |   ||        |    | |:|       :|  :  : | |  |
Db     83 VKV-----ETEDQVLATFCGRETTDTEQT--PGQEVVLSPGTFMSVTFRSDFSNEERFTG 135
```

```
Qy     327 FEATFFQLP----------------------------------------RMSSCGGRL 344
           |:| :  :                                          |:   |
Db     136 FDAHYMAVDVDECKEREDEELSCDHYCHNYIGGYYCSCRFGYILHTDNRTCRVECSGNLF 195

Qy     345 RKAQGTFNSPYYPGHYPPNIDCTWNIEVPNNQHVKVSFKFFYLLE--PGVPAGTCPKDYV 402
           :  ||  || ||  :|: :|::: |::     | : |:  : :|  | ||   || ||:
Db     196 TQRTGTITSPDYPNPYPKSSECSYTIDLEEGFMVSLQFEDIFDIEDHPEVP---CPYDYI 252

Qy     403 EIN-GEK----YCGERSQFVVTSNSNKITVRFHSDQSYTDTGFLAEYLS----------- 446
           :|  | |    :|||:|   ::: ::.:: | |4 |  : |:   | :
Db     253 KIKAGSKVWGPFCGEKSPEPISTQTHSVQILFRSDNSGENRGWRLSYRAAGNECPKLQPP 312

Qy     447 -YDSSDPCPGQFTCRTGRCIRKELRCDGWADCTDHSDELNCSCDAGH-----------Q 493
            |   :|    :: :                        |::  ||| |:           |
Db     313 VYGKFEPSQAVYSFK--------------------DQVLVSCDTGYKVLKDNGVMDTFQ 351

Qy     494 FTC------KNKF--CKPLFWVCDSVNDCGDNSDEQGCSCPAQTFRCSNGKCLSKSQQCN 545
           |       ||  ||       : ||| :  |      ||  |     ||:
Db     352 IECLKDGAWSNKIPTCK--------IVDCGAPA---GLKHGLVTFSTRNNLTTYKSE--- 397

Qy     546 GKDDCGDGSDEASCPK---------VNVVTCTKH-TYRCLNGLCLSKGNPECDGKEDCSD 595
                  || :          | ||: | |:      | :  | |
Db     398 ---------IRYSCQQPYYKMLHNTTGVYTCSAHGTW---TNKVLKRSLPTCLPV----- 440

Qy     596 GSDEKDCDCGLRSFTRQ--ARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVS 653
                 ||:  |:|:  :|:  |  | :|  ||   |  |     || ||:  ||:::
Db     441 --------CGVPKFSRKQISRIFNGRPAQKGTMPWIAMLSHLNGQPFCGGSLLGSNWVLT 492

Qy     654 AAHCYIDDRGFRYSDPTQWTAFL----------GLHDQSQRSAPGVQERRLKRIISHPFF 703
           ||||    :    :||  :::|          | |  :||    |  :||. || :
Db     493 AAHCL--HQSLDPEEPTLHSSYLLSPSDFKIIMGKH-WRRRSDEDEQHLHVKRTTLHPLY 549

Qy     704 NDFTFDYDIALLELEKPAEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALIL 763
           |  ||: |: |:|| :    :  | |:|||:      |  : |:|||  |:       |
Db     550 NPSTFENDLGLVELSESPRLNDFVMPVCLPEQPST--EGTMVIVSGWG-KQFLQRFPENL 606

Qy     764 QKGEIRVINQTTCENL---LPQQITPRMMCVGFLSGGVDSCQGDSGGPLSSVEAD-GRIF 819
           : || ::|  ||:     | :::|  |:| |.  || |:| |||||: : :|:  : :
Db     607 MEIEIPIVNSDTCQEAYTPLKKKVTKDMICAGEKEGGKDACAGDSGGPMVTKDAERDQWY 666

Qy     820 QAGVVSWGDGCAQRNKPGVYTRLPLFRDWIKENTGV 855
           ||||||: |.:::: |||: :   :|||:  |||
Db     667 LVGVVSWGEDCGKKDRYGVYSYIYPNKDWIQRITGV 702


RESULT  11
CRAR_HUMAN
ID   CRAR_HUMAN     STANDARD;      PRT;   699 AA.
AC   P48740;
DT   01-FEB-1996 (Rel. 33, Created)
DT    1-FEB-1996 (Rel. 33, Last sequence update)
DT    -AUG-2001 (Rel. 40, Last annotation update)
DE   COMPLEMENT-ACTIVATING COMPONENT OF RA-REACTIVE FACTOR PRECURSOR
DE   (EC 3.4.21.-) (RA-REACTIVE FACTOR SERINE PROTEASE P100) (RARF)
DE   (MANNAN-BINDING LECTIN SERINE PROTEASE 1) (MANNOSE-BINDING PROTEIN
DE   ASSOCIATED SERINE PROTEASE) (MASP-1).
GN   MASP1 OR CRARF OR CRARF1 OR PRSS5.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Liver;
RX   MEDLINE=94059062; PubMed=8240317;
RA   Takada F., Takayama Y., Hatsuse H., Kawakami M.;
RT   "A new member of the C1s family of complement proteins found in a
RT   bactericidal factor, Ra-reactive factor, in human serum.";
RL   Biochem. Biophys. Res. Commun. 196:1003-1009(1993).
RN   [2]
RP   SEQUENCE FROM N.A.
RC   TISSUE=Fetal liver;
RX   MEDLINE=94289349; PubMed=8018603;
RA   Sato T., Endo Y., Matsushita M., Fujita T.;
RT   "Molecular characterization of a novel serine protease involved in
RT   activation of the complement system by mannose-binding protein.";
RL   Int. Immunol. 6:665-669(1994).
CC   -!- FUNCTION: COMPONENT OF THE BACTERICIDAL RA-REACTIVE FACTOR RARF
CC       WHICH SPECIFICALLY BINDS TO RA AND R2 POLYSACCHARIDES EXPRESSED BY
CC       CERTAIN ENTEROBACTERIA. IT TRIGGERS THE ACTIVATION OF COMPLEMENT
CC       CASCADE BY ACTIVATING THE C4 AND C2 COMPONENTS. IT ACTIVATES THE
CC       C4 COMPONENT BY CLEAVING THE ALPHA-CHAIN OF C4.
CC   -!- SUBUNIT: RARF CONSISTS OF A COMPLEMENT-ACTIVATING COMPONENT
CC       (CRARF) AND A POLYSACCHARIDE-BINDING (MANNOSE-BINDING) COMPONENT.
CC       CRARF IS AN HETERODIMER OF A HEAVY (P70) AND A LIGHT CHAIN (P29)
CC       LINKED BY A DISULFIDE BOND.
CC   -!- DOMAIN: CRARF HAS A MODULE ORGANIZATION SIMILAR TO C1R AND C1S.
CC   -!- SIMILARITY: CONTAINS 2 SUSHI (SCR) DOMAINS.
CC   -!- SIMILARITY: CONTAINS 2 CUB DOMAINS.
CC   -!- SIMILARITY: CONTAINS 1 EGF-LIKE DOMAIN.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; D17525; BAA04477.1; -.
DR   EMBL; D28593; BAA05928.1; -.
DR   HSSP; P00736; 1APQ.
DR   MEROPS; S01.198; -.
DR   MIM; 600521; -.
DR   InterPro; IPR000152; Asx_hydroxyl.
DR   InterPro; IPR000859; CUB.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR000561; EGF-like.
DR   InterPro; IPR001881; EGF_Ca.
DR   InterPro; IPR000436; Sushi_SCR_CCP.
DR   InterPro; IPR001254; Trypsin.
DR   Pfam; PF00431; CUB; 2.
DR   Pfam; PF00084; sushi; 2.
DR   Pfam; PF00089; trypsin; 1.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   SMART; SM00032; CCP; 2.
DR   SMART; SM00042; CUB; 2.
DR   SMART; SM00179; EGF_CA; 1.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS00010; ASX_HYDROXYL; 1.
DR   PROSITE; PS01180; CUB; 2.
DR   PROSITE; PS01186; EGF_2; 1.
DR   PROSITE; PS01187; EGF_CA; 1.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Hydrolase; Complement pathway; Serine protease; Protease;
KW   Glycoprotein; Sushi; Repeat; Signal; EGF-like domain; Hydroxylation.
FT   SIGNAL       1    19       POTENTIAL.
FT   CHAIN       20   699       COMPLEMENT-ACTIVATING COMPONENT OF
FT                              RA-REACTIVE FACTOR.
FT   CHAIN       20   448       70 KDA CHAIN OF P100 (P70).
FT   CHAIN      449   699       29 KDA CHAIN OF P100 (P29).
FT   DOMAIN      20   138       CUB 1.
FT   DOMAIN     139   182       EGF-LIKE, CALCIUM-BINDING (POTENTIAL).
FT   DOMAIN     185   297       CUB 2.
FT   DOMAIN     300   363       SUSHI 1.
FT   DOMAIN     366   433       SUSHI 2.
FT   DOMAIN     449   699       SERINE PROTEASE.
FT   ACT_SITE   490   490       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE   552   552       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE   646   646       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   MOD_RES    159   159       HYDROXYLATION (POTENTIAL).
FT   DISULFID    73    91       POTENTIAL.
```

```
FT   DISULFID    143    157       POTENTIAL.
FT   DISULFID    153    166       POTENTIAL.
FT   DISULFID    168    181       POTENTIAL.
FT   DISULFID    185    212       POTENTIAL.
FT   DISULFID    242    260       POTENTIAL.
FT   DISULFID    301    349       POTENTIAL.
FT   DISULFID    329    362       POTENTIAL.
FT   DISULFID    367    414       POTENTIAL.
FT   DISULFID    397    432       POTENTIAL.
FT   DISULFID    436    572       INTERCHAIN (POTENTIAL).
FT   DISULFID    614    631       POTENTIAL.
FT   DISULFID    642    672       POTENTIAL.
FT   CARBOHYD     49     49       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    178    178       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    385    385       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    407    407       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT    235    235       E -> Q (IN REF. 2).
FT   CONFLICT    285    285       G -> A (IN REF. 2).
FT   CONFLICT    499    499       K -> G (IN REF. 2).
FT   CONFLICT    543    543       K -> Q (IN REF. 2).
FT   CONFLICT    643    643       S -> A (IN REF. 2).
SQ   SEQUENCE   699 AA;  79261 MW;  45DEC2EA6EB40151 CRC64;


  Query Match          12.3%;  Score 577;  DB 1;  Length 699;
  Best Local Similarity  25.4%;  Pred. No. 1.2e-31;
  Matches  194;  Conservative  110;  Mismatches  268;  Indels  192;  Gaps   32;

Qy    203 SKTVQRTQDNSCSFGLHARGVELMRFTTPGFPDSPYPAHARCQWALRGDADSVLSLTFRS 262
Db     15 SKASAHTVELNNMFG---------QIQSPGYPDS-YPSDSEVTWNITVPDGFRIKLYFMH  64

Qy    263 FDLASCDERGSDLVTVYNTLSPMEPHALVQLCG----------TYPPSY-NLTFH 306
Db     65 FNLESSYLCEYDYVKV-----ETEDQVLATFCGRETTDTEQTPGQEVVLSPGSFMSITFR 119

Qy    307 SSQNVLLITLITNTERRHPGFEATFFQLP------------------------------ 335
Db    120 SD---------FSNEERFTGFDAHYMAVDVDECKEREDEELSCDHYCHNYIGGYYCSCRF 170

Qy    336 -----------RMSSCGGRLRKAQGTFNSPYYPGHYPPNIDCTWNIEVPNNQHVKVSFKF 384
Db    171 GYILHTDNRTCRVECSDNLFTQRTGVITSPDFPNPYPKSSECLYTIELEEGFMVNLQFED 230

Qy    385 FYLLE--PGVPAGTCPKDYVEIN-GEK----YCGERSQFVVTSNSNKITVRFHSDQSYTD 437
Db    231 IFDIEDHPEVP---CPYDYIKIKVGPKVLGPFCGEKAPEPISTQSHSVLILFHSDNSGEN 287

Qy    438 TGFLAEYLSYDSSDPCP-------GQFTCRTGRCIRKELRCDGWADCTDHSDELNCSCDA 490
Db    288 RGWRLSYRA--AGNECPELQPPVHGKIEPSQAKYFFK--------------DQVLVSCDT 331

Qy    491 GHQF--------TCKNKFCKPLFW-----VCDSVNDCGDNSDEQGCSCPAQ------TFR 531
Db    332 GYKVLKDNVEMDTFQIECLKDGTWSNKIPTCKIVD----------CRAPGELEHGLITFS 381

Qy    532 CSNGKCLSKSQQCNGKDDCGDGSDEASCPKVNVVTCTKHTYRC------LNGLCLSKGNP 585
Db    382 TRNNLTTYKSEI---KYSCQE------PYYKMLNNNTGIYTCSAQGVWMNKV-LGRSLP 430

Qy    586 ECDGKEDCSDGSDEKDCDCGLRSFTRQ--ARVVGGTDADEGEWPWQVSLHALGQGHICGA 643
Db    431 TCLPV------------CGLPKFSRKLMARIFNGRPAQKGTTPWIAMLSHLNGQPFCGG 477

Qy    644 SLISPNWLVSAAHCY---ID--DRGFRYSD---PTQWTAFLGLHDQSQRSAPGVQERRLK 695
Db    478 SLLGSSWIVTAAHCLHQSLDPKDPTLRDSDLLSPSDFKIILGKH-WRLRSDENEQHLGVK 536

Qy    696 RIISHPFFNDFTFDYDIALLELEKPAEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQY 755
Db    537 HTTLHPKYDPNTFEND VALVELLESPVLNAFVMPICLPEGPQ--QEGAMVIVSGWG-KQF 593

Qy    756 GGTGALILQKGEIRVINQTTCENL---LPQQITPRMMCVGFLSGGVDSCQGDSGGPLSSV 812
```

```
Db    594 LQRFPETLMEIEIPIVDHSTCQKAYAPLKKKVTRDMICAGEKEGGKDACSGDSGGPMVTL 653

Qy    813 EAD-GRIFQAGVVSWGDGCAQRNKPGVYTRLPLFRDWIKENTGV 855
Db    654 NRERCQWYLVGTVSWGDDCGKKDRYGVYSYIHHNKDWIQRVTGV 697


RESULT  12
TMS2_HUMAN
ID   TMS2_HUMAN     STANDARD;      PRT;   492 AA.
AC   O15393;
DT   15-JUL-1998 (Rel. 36, Created)
DT   15-JUL-1998 (Rel. 36, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   TRANSMEMBRANE PROTEASE, SERINE 2 (EC 3.4.21.-).
GN   TMPRSS2.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=97468144; PubMed=9325052;
RA   Paoloni-Giacobino A., Chen H., Peitsch M.C., Rossier C.,
RA   Antonarakis S.E.;
RT   "Cloning of the TMPRSS2 gene, which encodes a novel serine protease
RT   with transmembrane, LDLRA, and SRCR domains and maps to 21q22.3.";
RL   Genomics 44:309-320(1997).
RN   [2]
RP   SEQUENCE FROM N.A.
RX   PubMed=11414763;
RA   Teng D.H., Chen Y., Lian L., Ha P.C., Tavtigian S.V., Wong A.K.;
RT   "Mutation analyses of 268 candidate genes in human tumor cell lines.";
RL   Genomics 74:352-364(2001).
RN   [3]
RP   TISSUE SPECIFICITY.
RX   MEDLINE=21104370; PubMed=11169526;
RA   Vaarala M.H., Porvari K.S., Kellokumpu S., Kyllonen A.P., Vihko P.T.;
RT   "Expression of transmembrane serine protease TMPRSS2 in mouse and
RT   human tissues.";
RL   J. Pathol. 193:134-140(2001).
CC   -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN.
CC   -!- TISSUE SPECIFICITY: EXPRESSED STRONGLY IN SMALL INTESTINE. ALSO
CC       EXPRESSED IN PROSTATE, COLON, STOMACH, AND SALIVARY GLAND.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   -!- SIMILARITY: CONTAINS 1 SRCR DOMAIN.
CC   -!- SIMILARITY: CONTAINS 1 LDL-RECEPTOR CLASS A DOMAIN.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; U75329; AAC51784.1; -.
DR   EMBL; AF123453; AAD37117.1; -.
DR   HSSP; P00763; 1DPO.
DR   MIM; 602060; -.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR002172; LDL_recept_A.
DR   InterPro; IPR001190; SRCR.
DR   InterPro; IPR001254; Trypsin.
DR   Pfam; PF00057; ldl_recept_a; 1.
DR   Pfam; PF00089; trypsin; 1.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   SMART; SM00192; LDLa; 1.
DR   SMART; SM00202; SR; 1.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS01209; LDLRA_1; 1.
```

```
DR   PROSITE; PS50068; LDLRA_2; 1.
DR   PROSITE; PS00420; SRCR_1; FALSE_NEG.
DR   PROSITE; PS50287; SRCR_2; 1.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Hydrolase; Serine protease; Transmembrane; Signal-anchor.
FT   DOMAIN        1     84       CYTOPLASMIC (POTENTIAL).
FT   TRANSMEM     85    105       SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN)
FT                                 (POTENTIAL).
FT   DOMAIN      106    492       EXTRACELLULAR (POTENTIAL).
FT   DOMAIN      112    149       LDL-RECEPTOR CLASS A.
FT   DOMAIN      150    242       SRCR.
FT   DOMAIN      256    492       SERINE PROTEASE.
FT   ACT_SITE    296    296       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    345    345       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    441    441       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   SITE        255    256       CLEAVAGE (POTENTIAL).
FT   DISULFID    113    126       BY SIMILARITY.
FT   DISULFID    120    139       BY SIMILARITY.
FT   DISULFID    133    148       BY SIMILARITY.
FT   DISULFID    172    231       BY SIMILARITY.
FT   DISULFID    185    241       BY SIMILARITY.
FT   DISULFID    244    365       BY SIMILARITY.
FT   DISULFID    281    297       BY SIMILARITY.
FT   DISULFID    410    426       BY SIMILARITY.
FT   DISULFID    437    465       BY SIMILARITY.
FT   CARBOHYD    213    213       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    249    249       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CONFLICT    242    242       L -> I (IN REF. 2).
FT   CONFLICT    329    329       Q -> E (IN REF. 2).
FT   CONFLICT    449    449       N -> K (IN REF. 2).
FT   CONFLICT    489    491       KAN -> RAD (IN REF. 2).
SQ   SEQUENCE   492 AA;  53847 MW;  7EAAFFDA18609DDA CRC64;
```

```
  Query Match             12.0%;  Score 560.5;  DB 1;  Length 492;
  Best Local Similarity   33.9%;  Pred. No. 1e-30;
  Matches  131;  Conservative   58;  Mismatches  128;  Indels   69;  Gaps   14;

Qy     522 GCSCPAQTFRC-SNGKCLSKSQQCNGKDDCGDGSDEASCPKV---NVV------------ 565
           |  |      | |:| |::  |  |:|   |  | ||  | ::   | :
Db     110 GSKCSNSGIECDSSGTCINPSNWCDGVSHCPGGEDENRCVRLYGPNFILQMYSSQRKSWH 169

Qy     566 --------------TCT----KHTYRCLNG------------LCLSKGNPECDGKEDCSD 595
                         |    |:  :   |             |  | || :   |   ||
Db     170 PVCQDDWNENYGRAACRDMGYKNNFYSSQGIVDDSGSTSFMKLNTSAGNVDIYKKLYHSD 229

Qy     596 GSDEK--------DCDCGLRSFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLIS 647
              |         |  | | :||:|:|||  |  | ||||||||  :   |:|| |:|:
Db     230 ACSSKAVVSLRCLACGVNLNS-SRQSRIVGGESALPGAWPWQVSLH-VQNVHVCGGSIIT 287

Qy     648 PNWLVSAAHCYIDDRGFRYSDPTQWTAFLGLHDQS-QRSAPGVQERRLKRIISHPFFNDF 706
           | |:|:|||||         ::|  |||| |:  ||      | |   ::::|||| ::
Db     288 PEWIVTAAHCVEKP----LNNPWHWTAFAGILRQSFMFYGAGYQ---VQKVISHPNYDSK 340

Qy     707 TFDYDIALLELEKPAEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKG 766
           | : ||||::|:||  :: :|:|:|||:   :  :  |::||| |:  |  : :|
Db     341 TKNNDIALMKLQKPLTFNDLVKPVCLPNPGMMLQPEQLCWISGWGATEEKGKTSEVLNAA 400

Qy     767 EIRVINQTTCEN--LLPQQITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRI-FQAGV 823
           :: :|    | :  :    ||| |:| ||| | ||||||||||||  | ::  | :  |
Db     401 KVLLIETQRCNSRYVYDNLITPAMICAGFLQGNVDSCQGDSGGPL--VTSNNNIWWLIGD 458

Qy     824 VSWGDGCAQRNKPGVYTRLPLFRDWI 849
           ||| |||:  :||||  : :| |||
Db     459 TSWGSGCAKAYRPGVYGNVMVFTDWI 484
```

```
RESULT  13
KAL_MOUSE
ID   KAL_MOUSE       STANDARD;       PRT;   638 AA.
AC   P26262;
DT   01-MAY-1992 (Rel. 22, Created)
DT   01-MAY-1992 (Rel. 22, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   PLASMA KALLIKREIN PRECURSOR (EC 3.4.21.34) (PLASMA PREKALLIKREIN)
DE   (KININOGENIN) (FLETCHER FACTOR).
GN   KLK3 OR PK.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A., AND PARTIAL SEQUENCE.
RC   STRAIN=BALB/C; TISSUE=Liver;
RX   MEDLINE=91090844; PubMed=2264928;
RA   Seidah N.G., Sawyer N., Hamelin J., Mion P., Beaubien G.,
RA   Brachpapa L., Rochemont J., Mbikay M., Chretien M.;
RT   "Mouse plasma kallikrein: cDNA structure, enzyme characterization,
RT   and comparison of protein and mRNA levels among species.";
RL   DNA Cell Biol. 9:737-748(1990).
CC   -!- FUNCTION: THE ENZYME CLEAVES LYS-ARG AND ARG-SER BONDS. IT
CC       ACTIVATES, IN A RECIPROCAL REACTION, FACTOR XII AFTER ITS BINDING
CC       TO A NEGATIVELY CHARGED SURFACE. IT ALSO RELEASES BRADYKININ FROM
CC       HMW KININOGEN AND MAY ALSO PLAY A ROLE IN THE RENIN-ANGIOTENSIN
CC       SYSTEM BY CONVERTING PRORENIN INTO RENIN.
CC   -!- SUBUNIT: THE ZYMOGEN IS ACTIVATED BY FACTOR XIIA, WHICH CLEAVES
CC       THE MOLECULE INTO A LIGHT CHAIN, WHICH CONTAINS THE ACTIVE SITE,
CC       AND A HEAVY CHAIN, WHICH ASSOCIATES WITH HMW KININOGEN. THESE
CC       CHAINS ARE LINKED BY ONE OR MORE DISULFIDE BONDS.
CC   -!- SIMILARITY: CONTAINS 4 APPLE DOMAINS.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY. PLASMA KALLIKREIN SUBFAMILY.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; M58588; AAA63393.1; -.
DR   PIR; A36557; KQMSPL.
DR   HSSP; P00750; 1RTF.
DR   MEROPS; S01.212; -.
DR   MGD; MGI:102849; Klk3.
DR   InterPro; IPR000177; Apple.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR003014; PAN.
DR   InterPro; IPR001254; Trypsin.
DR   Pfam; PF00024; PAN; 4.
DR   Pfam; PF00089; trypsin; 1.
DR   PRINTS; PR00005; APPLEDOMAIN.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   SMART; SM00223; APPLE; 4.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS00495; APPLE; 4.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Hydrolase; Serine protease; Glycoprotein; Plasma; Zymogen; Signal;
KW   Fibrinolysis; Blood coagulation; Inflammatory response; Liver;
KW   Repeat.
FT   SIGNAL        1     19
FT   CHAIN .      20    390       PLASMA KALLIKREIN HEAVY CHAIN.
FT   CHAIN       391    638       PLASMA KALLIKREIN LIGHT CHAIN.
FT   DOMAIN       20    105       APPLE 1.
FT   DOMAIN      110    195       APPLE 2.
FT   DOMAIN      200    285       APPLE 3.
FT   DOMAIN      291    376       APPLE 4.
FT   DOMAIN      389    621       SERINE PROTEASE.
FT   CARBOHYD    127    127       N-LINKED (GLCNAC. . .) (PROBABLE).
FT   CARBOHYD    308    308       N-LINKED (GLCNAC. . .) (PROBABLE).
FT   CARBOHYD    396    396       N-LINKED (GLCNAC. . .) (PROBABLE).
```

```
FT   CARBOHYD    453    453       O-LINKED (PROBABLE).
FT   CARBOHYD    494    494       N-LINKED (GLCNAC. . .) (PROBABLE).
FT   ACT_SITE    434    434       CHARGE RELAY SYSTEM.
FT   ACT_SITE    483    483       CHARGE RELAY SYSTEM.
FT   ACT_SITE    578    578       CHARGE RELAY SYSTEM.
FT   DISULFID     21    104       BY SIMILARITY.
FT   DISULFID     47     77       BY SIMILARITY.
FT   DISULFID     51     57       BY SIMILARITY.
FT   DISULFID    111    194       BY SIMILARITY.
FT   DISULFID    137    166       BY SIMILARITY.
FT   DISULFID    141    147       BY SIMILARITY.
FT   DISULFID    201    284       BY SIMILARITY.
FT   DISULFID    227    256       BY SIMILARITY.
FT   DISULFID    231    237       BY SIMILARITY.
FT   DISULFID    292    375       BY SIMILARITY.
FT   DISULFID    318    347       BY SIMILARITY.
FT   DISULFID    322    328       BY SIMILARITY.
FT   DISULFID    340    345       BY SIMILARITY.
FT   DISULFID    383    503       BY SIMILARITY.
FT   DISULFID    419    465       BY SIMILARITY.
FT   DISULFID    517    584       BY SIMILARITY.
FT   DISULFID    548    563       BY SIMILARITY.
FT   DISULFID    574    602       BY SIMILARITY.
SQ   SEQUENCE   638 AA;  71368 MW;  CC27C93AB1086599 CRC64;


  Query Match             11.4%;  Score 533.5;  DB 1;  Length 638;
  Best Local Similarity   27.1%;  Pred. No. 9.2e-29;
  Matches  183;  Conservative   86;  Mismatches  233;  Indels  173;  Gaps   31;

Qy    271 RGSDLVTVYNTLSPMEPHALVQLCGTYPPSYNLTFHSSQNVLLITLITNTERRHPGF--E 328
Db     32 RGGDLAAIYTP----DAQYCQKMCTFHPRCLLFSF------LAVTPPKETNKRFGCFMKE 81

Qy    329 ATFFQLPRMSSCG---GRLRKAQGTFNSPYYPGHYPP--------NIDCTWNIE-----V 372
Db     82 SITGTLPRIHRTGAISGHSLKQCGHQISACHRDIYKGLDMRGSNFNISKTDNIEECQKLC 141

Qy    373 PNNQHVK---VSFKFFYLLE-----------PGVPAGT---------------------C 397
Db    142 TNNFHCQFFTYATSAFYRPEYRKKCLLKHSASGTPTSIKSADNLVSGFSLKSCALSEIGC 201

Qy    398 PKD------YVEINGEKCGERS---QFVVTSNSNKITVRFHSDQSYTDTG----FLAEY 444
Db    202 PMDIFQHSAFADLNVSQVITPDAFVCRTICTFHPNCLFFTFYTNEWETESQRNVCFLKTS 261

Qy    445 LSYDSSDPCPGQ--------FTCRTGR---CIRKELRCDGWADCTDHSDELNCSCDAGH- 492
Db    262 KSGRPSPPIPQENAISGYSLLTCRKTRPEPCHSKI-----YSGVDFEGEELNVTFVQGAD 316

Qy    493 --QFTCKNKFCKPLFWVCDSV-NDCGDNSDEQGCSCPAQTFRCSNGKCLSKSQQCNGKDD 549
Db    317 VCQETC-TKTIRCQFFIYSLLPQDC----KEEGCKC---SLRLST--------------- 353

Qy    550 CGDGSDEASCPKVNVVTCTKHTYRCLNG-------LCLSKGNPECDGKEDCSDGSDEKDCD 603
Db    354 --DGSP------------TRITYGMQGSSGYSLRLCKLVDSPDCTTKIN---------- 388

Qy    604 CGLRSFTRQARVVGGTDADEGEWPWQVSLHA--LGQGHICGASLISPNWLVSAAHCYIDD 661
Db    389 ---------ARIVGGTNASLGEWPWQVSLQVKLVSQTHLCGGSIIGRQWVLTAAHCF--- 436

Qy    662 RGFRYSDPTQWTAFLGLHDQSQ--RSAPGVQERRLKRIISHPFFNDFTFDYDIALLELEK 719
Db    437 DGIPY--PDVWRIYGGILSLSEITKETP---SSRIKELIIHQEYKVSEGNYDIALIKLQT 491

Qy    720 PAEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTTCENL 779
Db    492 PLNYTEFQKPICLPSKADTNTIYTNCWVTGWGYTKEQGETQNILQKATIPLVPNEECQKK 551

Qy    780 LPQQ-ITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRNKPGV 838
Db    552 YRDYVINKQMICAGYKEGGTDACKGDSGGPL-VCKHSGRWQLVGITSWGEGCGRKDQPGV 610

Qy    839 YTRLPLFRDWIKENT 853
Db    611 YTKVSEYMDWILEKT 625


RESULT  14
TMS3_HUMAN
ID   TMS3_HUMAN     STANDARD;      PRT;   454 AA.
AC   P57727;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   TRANSMEMBRANE PROTEASE, SERINE 3 (EC 3.4.21.-) (SERINE PROTEASE
DE   TADG-12) (TUMOR ASSOCIATED DIFFERENTIALLY-EXPRESSED GENE-12 PROTEIN).
GN   TMPRSS3 OR TADG12 OR ECHOS1.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A. (ISOFORMS A AND TRUNCATED).
RC   TISSUE=Ovarian carcinoma;
RX   MEDLINE=20521358; PubMed=11068177;
RA   Underwood L.J., Shigemasa K., Tanimoto H., Beard J.B., Schneider E.N.,
RA   Wang Y., Parmley T.H., O'Brien T.J.;
RT   "Ovarian tumor cells express a novel multi-domain cell surface serine
RT   protease.";
RL   Biochim. Biophys. Acta 1502:337-350(2000).
RN   [2]
RP   SEQUENCE FROM N.A. (ISOFORMS A; B/C AND D), AND VARIANT ILE-53.
RX   MEDLINE=20578749; PubMed=11137999;
RA   Scott H.S., Kudoh J., Wattenhofer M., Shibuya K., Berry A., Chrast R.,
RA   Guipponi M., Wang J., Kawasaki K., Asakawa S., Minoshima S.,
RA   Younus F., Mehdi S.Q., Radhakrishna U., Papasavvas M.P., Gehrig C.,
RA   Rossier C., Korostishevsky M., Gal A., Shimizu N., Bonne-Tamir B.,
RA   Antonarakis S.E.;
RT   "Insertion of beta-satellite repeats identifies a transmembrane
RT   protease causing both congenital and childhood onset autosomal
RT   recessive deafness.";
RL   Nat. Genet. 27:59-63(2001).
CC   -!- FUNCTION: PROBABLE PROTEASE.
CC   -!- SUBCELLULAR LOCATION: TYPE II MEMBRANE PROTEIN (POTENTIAL).
CC   -!- ALTERNATIVE PRODUCTS: 4 ISOFORMS; A (SHOWN HERE), B/C, D AND
CC       TRUNCATED/TADG-12V; ARE PRODUCED BY ALTERNATIVE SPLICING.
CC   -!- TISSUE SPECIFICITY: EXPRESSED IN MANY TISSUES. THE TRUNCATED
CC       ISOFORM IS FOUND AT INCREASED LEVELS IN SOME CARCINOMAS.
CC   -!- DISEASE: DEFECTS IN TMPRSS3 ARE A CAUSE OF TWO FORMS OF AUTOSOMAL
CC       NEUROSENSORY CHILDHOOD-ONSET FORMS OF DEAFNESS, DFNB8 AND DFNB10.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   -!- SIMILARITY: CONTAINS 1 SRCR DOMAIN.
CC   -!- SIMILARITY: CONTAINS 1 LDL-RECEPTOR CLASS A DOMAIN.
CC   ---------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ---------------------------------------------------------------------
DR   EMBL; AF201380; AAG37012.1; -.
DR   EMBL; AB038157; BAB20077.1; -.
DR   EMBL; AB038158; BAB20078.1; -.
DR   EMBL; AB038159; BAB20079.1; -.
DR   EMBL; AB038160; BAB20080.1; -.
DR   MIM; 605511; -.
DR   MIM; 601072; -.
DR   MIM; 605316; -.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR002172; LDL_recept_A.
DR   InterPro; IPR001190; SRCR.
```

```
DR   InterPro; IPR001254; Trypsin.
DR   Pfam; PF00057; ldl_recept_a; 1.
DR   Pfam; PF00530; SRCR; 1.
DR   Pfam; PF00089; trypsin; 1.
DR   SMART; SM00192; LDLa; 1.
DR   SMART; SM00202; SR; 1.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS01209; LDLRA_1; 1.
DR   PROSITE; PS50068; LDLRA_2; 1.
DR   PROSITE; PS00420; SRCR_1; FALSE_NEG.
DR   PROSITE; PS50287; SRCR_2; 1.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Hydrolase; Serine protease; Transmembrane; Signal-anchor; Deafness;
KW   Alternative splicing; Polymorphism.
FT   DOMAIN        1     48       CYTOPLASMIC (POTENTIAL).
FT   TRANSMEM      49    69       SIGNAL-ANCHOR (TYPE-II MEMBRANE PROTEIN)
FT                                (POTENTIAL).
FT   DOMAIN        70    454      EXTRACELLULAR (POTENTIAL).
FT   DOMAIN        72    108      LDL-RECEPTOR CLASS A.
FT   DOMAIN        109   205      SRCR.
FT   DOMAIN        217   454      SERINE PROTEASE.
FT   ACT_SITE      257   257      CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE      304   304      CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE      401   401      CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   SITE          216   217      CLEAVAGE (POTENTIAL).
FT   DISULFID      73    85       BY SIMILARITY.
FT   DISULFID      79    98       BY SIMILARITY.
FT   DISULFID      92    107      BY SIMILARITY.
FT   DISULFID      129   194      BY SIMILARITY.
FT   DISULFID      142   204      BY SIMILARITY.
FT   DISULFID      207   324      BY SIMILARITY.
FT   DISULFID      242   258      BY SIMILARITY.
FT   DISULFID      338   407      BY SIMILARITY.
FT   DISULFID      370   386      BY SIMILARITY.
FT   DISULFID      397   425      BY SIMILARITY.
FT   CARBOHYD      221   221      N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   VARSPLIC      1     127      MISSING (IN ISOFORM B/C).
FT   VARSPLIC      318   454      EMIQPVCLPNSEENFPDGKVCWTSGWGATEDGAGDASPVLN
FT                                HAAVPLISNKICNHRDVYGGIISPSMLCAGYLTGGVDSCQG
FT                                DSGGPLVCQERRLWKLVGATSFGIGCAEVNKPGVYTRVTSF
FT                                LDWIHEQMERDLKT -> GTSGSLCGSAALPLFQEDLQLLI
FT                                EAFL (IN ISOFORM D).
FT   VARSPLIC      261   293      DLYLPKSWTIQVGLVSLLDNPAPSHLVEKIVYH -> EIVA
FT                                PRERADRRGRKLLCWRKPTKMKGPRPSHS (IN
FT                                TRUNCATED ISOFORM).
FT   VARSPLIC      294   454      MISSING (IN TRUNCATED ISOFORM).
FT   VARIANT       53    53       V -> I.
FT                                /FTId=VAR_010781.
FT   CONFLICT      46    54       LKFFPIIVI -> FEVFSQSSSL (IN REF. 1).
FT   CONFLICT      90    90       A -> T (IN REF. 1).
FT   CONFLICT      350   350      MISSING (IN REF. 1).
FT   CONFLICT      369   395      ICNHRDVYGGIISPSMLCAGYLTGGVD -> DLQPQGRVRW
FT                                HHLPLHALRGLPDGWRWN (IN REF. 1).
FT   CONFLICT      427   427      E -> D (IN REF. 1).
SQ   SEQUENCE   454 AA;  49404 MW;  57ECC3678F7D6AFF CRC64;

  Query Match           11.4%;  Score 533;  DB 1;  Length 454;
  Best Local Similarity  33.5%;  Pred. No. 6.6e-29;
  Matches  128;  Conservative   54;  Mismatches  128;  Indels   72;  Gaps   15;

Qy    530 FRC-SNGKCLSKSQQCNGKDDCGDGSDEASCPKV-------NVVTCTKHTYRC------- 574
          :|| |: ||:    :|:|  || || ||  | :|       | |      |
Db     77 YRCRSSFKCIELIARCDGVSDCKDGEDEYRCVRVGGQNAVLQVFTAASWKTMCSDDWKGH 136

Qy    575 -LNGLCLSKGNP----------ECDG-------------------------KEDCSD 595
           |  |  | ||           :|                         :| |:
Db    137 YANVACAQLGFPSYVSSDNLRVSSLEGQFREEFVSIDHLLPDDKVTALHHSVYVREGCAS 196

Qy    596 GS-DEKDCD-CGLRSFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVS 653
          |      |  || |    :|:|||  :   :|||| ||   |  |:|| |:|:| |:::

Db    197 GHVVTLQCTACGHRR-GYSSRIVGGNMSLLSQWPWQASLQFQGY-HLCGGSVITPLWIIT 254

Qy    654 AAHCYIDDRGFRYSDPTQWTAFLGLHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIA 713
          ||||  |        |  ||  :||       ||      :::|: |  :       |||
Db    255 AAHCVYD-----LYLPKSWTIQVGLVSLLDNPAP---SHLVEKIVYHSKYKPKRLGNDIA 306

Qy    714 LLELEKPAEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYG-GTGALILQKGEIRVIN 772
          |:::  |  :: |:::|:|||::   || ||  | :||| |: | |  : :|    : :|:
Db    307 LMKLAGPLTFNEMIQPVCLPNSEENFPDGKVCWTSGWGATEDGAGDASPVLNHAAVPLIS 366

Qy    773 QTTC--ENLLPQQITPRMMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQ-AGVVSWGDG 829
              |   ::    |:| |:| |:|:||||||||||||||  |  : |:::  |  |:| |
Db    367 NKICNHRDVYGGIISPSMLCAGYLTGGVDSCQGDSGGPL--VCQERRLWKLVGATSFGIG 424

Qy    830 CAQRNKPGVYTRLPLFRDWIKE 851
          ||: ||||||||:  | ||| |
Db    425 CAEVNKPGVYTRVTSFLDWIHE 446


RESULT  15
CFAI_MOUSE
ID   CFAI_MOUSE     STANDARD;      PRT;   603 AA.
AC   Q61129; Q9WU07;
DT   20-AUG-2001 (Rel. 40, Created)
DT   20-AUG-2001 (Rel. 40, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   COMPLEMENT FACTOR I PRECURSOR (EC 3.4.21.45) (C3B/C4B INACTIVATOR).
GN   IF OR CFI.
OS   Mus musculus (Mouse).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
OX   NCBI_TaxID=10090;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=BALB/C;
RX   MEDLINE=96175003; PubMed=8604219;
RA   Minta J.O., Wong M.J., Kozak C.A., Kunnath-Muglia L.M., Goldberger G.;
RT   "cDNA cloning, sequencing and chromosomal assignment of the gene for
RT   mouse complement factor I (C3b/C4b inactivator): identification of a
RT   species specific divergent segment in factor I.";
RL   Mol. Immunol. 33:101-112(1996).
RN   [2]
RP   SEQUENCE OF 114-334 FROM N.A.
RC   STRAIN=129/SV; TISSUE=Kidney;
RX   MEDLINE=99220287; PubMed=10204086;
RA   Yun Y.-S., Goldberger G., Minta J.O.;
RT   "Cloning and characterization of the non-catalytic heavy chain of
RT   mouse complement factor I gene: structure comparison with the human
RT   homologue.";
RL   Biochem. Mol. Biol. Int. 47:493-500(1999).
CC   -!- FUNCTION: RESPONSIBLE FOR CLEAVING THE ALPHA-CHAINS OF C4B AND C3B
CC       IN THE PRESENCE OF THE COFACTORS C4-BINDING PROTEIN AND FACTOR H
CC       RESPECTIVELY.
CC   -!- SUBUNIT: HETERODIMER OF A LIGHT AND HEAVY CHAINS LINKED BY
CC       DISULFIDE BONDS.
CC   -!- SUBCELLULAR LOCATION: EXTRACELLULAR.
CC   -!- TISSUE SPECIFICITY: PLASMA.
CC   -!- SIMILARITY: BELONGS TO PEPTIDASE FAMILY S1; ALSO KNOWN AS THE
CC       TRYPSIN FAMILY.
CC   -!- SIMILARITY: CONTAINS 1 KAZAL-LIKE DOMAIN.
CC   -!- SIMILARITY: CONTAINS 2 LDL-RECEPTOR CLASS A DOMAINS.
CC   -!- SIMILARITY: CONTAINS 1 SRCR DOMAIN.
CC   ----------------------------------------------------------------------
CC   This SWISS-PROT entry is copyright. It is produced through a collaboration
CC   between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC   the European Bioinformatics Institute. There are no restrictions on its
CC   use by non-profit institutions as long as its content is in no way
CC   modified and this statement is not removed. Usage by and for commercial
CC   entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC   or send an email to license@isb-sib.ch).
CC   ----------------------------------------------------------------------
DR   EMBL; U47810; AAB00438.1; -.
DR   EMBL; AF100565; AAD32965.1; -.
```

```
DR   EMBL; AF100559; AAD32965.1; JOINED.
DR   EMBL; AF100560; AAD32965.1; JOINED.
DR   EMBL; AF100561; AAD32965.1; JOINED.
DR   EMBL; AF100562; AAD32965.1; JOINED.
DR   EMBL; AF100563; AAD32965.1; JOINED.
DR   EMBL; AF100564; AAD32965.1; JOINED.
DR   HSSP; P00750; 1RTF.
DR   MEROPS; S01.199; -.
DR   MGD; MGI:105937; Cfi.
DR   InterPro; IPR001190; SRCR.
DR   InterPro; IPR001254; Trypsin.
DR   InterPro; IPR001314; Chymotrypsin.
DR   InterPro; IPR002172; LDL_recept_A.
DR   InterPro; IPR002350; kazal.
DR   InterPro; IPR003884; Fact_MAC.
DR   Pfam; PF00050; kazal; 1.
DR   Pfam; PF00057; ldl_recept_a; 2.
DR    fam; PF00530; SRCR; 1.
DR   fam; PF00089; trypsin; 1.
DR   PRINTS; PR00722; CHYMOTRYPSIN.
DR   SMART; SM00057; FIMAC; 1.
DR   SMART; SM00280; KAZAL; 1.
DR   SMART; SM00192; LDLa; 2.
DR   SMART; SM00202; SR; 1.
DR   SMART; SM00020; Tryp_SPc; 1.
DR   PROSITE; PS01209; LDLRA_1; 1.
DR   PROSITE; PS50068; LDLRA_2; 2.
DR   PROSITE; PS00420; SRCR_1; FALSE_NEG.
DR   PROSITE; PS50287; SRCR_2; 1.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Complement pathway; Plasma; Glycoprotein; Hydrolase; Serine protease;
KW   Repeat; Signal.
FT   SIGNAL        1     18       POTENTIAL.
FT   CHAIN        19    603       COMPLEMENT FACTOR I.
FT   CHAIN        19    356       COMPLEMENT FACTOR I HEAVY CHAIN.
FT   CHAIN       361    603       COMPLEMENT FACTOR I LIGHT CHAIN.
FT   DOMAIN       63    109       KAZAL-LIKE.
FT   DOMAIN      117    217       SRCR.
FT   DOMAIN      218    262       LDL-RECEPTOR CLASS A 1.
FT   DOMAIN      263    299       LDL-RECEPTOR CLASS A 2.
FT   DOMAIN      361    603       SERINE PROTEASE.
FT   DISULFID    234    252       BY SIMILARITY.
FT   DISULFID    246      ?       BY SIMILARITY.
F    DISULFID    264    276       BY SIMILARITY.
F    DISULFID    271    289       BY SIMILARITY.
FT   DISULFID    283    298       BY SIMILARITY.
FT   DISULFID    386    402       BY SIMILARITY.
FT   DISULFID    487    551       BY SIMILARITY.
FT   DISULFID    515    530       BY SIMILARITY.
FT   DISULFID    541    570       BY SIMILARITY.
FT   CARBOHYD    106    106       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    116    116       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    174    174       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    182    182       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    267    267       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    514    514       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   CARBOHYD    556    556       N-LINKED (GLCNAC. . .) (POTENTIAL).
FT   ACT_SITE    401    401       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    449    449       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   ACT_SITE    545    545       CHARGE RELAY SYSTEM (BY SIMILARITY).
FT   CONFLICT    114    114       N -> K (IN REF. 2).
FT   CONFLICT    236    252       NGKHIPQEKACNGVNDC -> MGSTFLRRKPATVSMTV
FT                                (IN REF. 1).
SQ   SEQUENCE   603 AA;  67234 MW;  E2C1D43261C719E8 CRC64;


  Query Match           11.1%;  Score 518;  DB 1;  Length 603;
  Best Local Similarity  33.4%;  Pred. No. 9.5e-28;
  Matches  127;  Conservative   56;  Mismatches  161;  Indels   36;  Gaps   14;

Qy    494 FTCKNKFCKPLFWVCDSVNDCGDNSDEQGC-SCPAQTFRCSNGKCLSKSQQCNGKDDCGD 552
          | | |    |    |: |||||| |||  |  |      | :| |:    :|||: ||
Db    232 FQCVNGKHIPQEKACNGVNDCGDQSDELCCKGCRGNASLCKSGVCIPDQYKCNGEVDCIT 291

Qy    553 GSDEASCPKVNVVTCTKHTYRCLNGLCLSKGNPECDGKEDCSDGSDEK---------DCD 603
          | ||: | :       |   |   |   :   | :  |  : | | :
Db    292 GEDESRCEEDRQQNIPKGLARSAQG----EAEIETEETEMLTPGMDNERKRIKSLLPKLS 347

Qy    604 CGLR--SFTRQARVVGGTDADEGEWPWQVSLHALGQGHICGASLISPNWLVSAAHCYIDD 661
          ||::  : ||: ||:||  |: |::||||::   ||   ||   |   |:::||||
Db    348 CGVKRNTHTRRKRVIGGKPANVGDYPWQVAIKD-GQRITCGGIYIGGCWILTAAHCVRPS 406

Qy    662 RGFRYSDPTQWTAFLG-LHDQSQRSAPGVQERRLKRIISHPFFNDFTFDYDIALLEL--- 717
          |   |     ||| |  |   ||    |:|   :||:| |  :|  ||  ||||:|:
Db    407 RAHSYQ---VWTALLDWLKPNSQL---GIQ--TVKRVIVHEKYNGATFQNDIALIEMKMH 458

Qy    718 --EKPAEYSSMVRPICLPDASHVFPAGKAIWVTGWGHTQYGGTGALILQKGEIRVINQTT 775
            :|  |  : | | |:| : ::|      ::|||  :          |: ||: :|
Db    459 TGKKECELPNSV-PACVPWSPYLFQPNDRCIISGWGRGK-DNQKVYSLRWGEVDLIG--N 514

Qy    776 CENLLPQQITPR-MMCVGFLSGGVDSCQGDSGGPLSSVEADGRIFQAGVVSWGDGCAQRN 834
          |    | :   : | | |   | :|:|:|||||||   : :   :  |:||||: | :
Db    515 CSQFYPDRYYEKEMQCAGTRDGSIDACKGDSGGPLVCEDTNNVTYVWGIVSWGENCGKPE 574

Qy    835 KPGVYTRLPLFRDWIKENTG 854
          ||||||:  : |||  : |
Db    575 FPGVYTRVANYFDWISYHVG 594


Search completed: January 20, 2002, 09:37:13
Job time: 144 sec
```